5-2013

DEVELOPMENT AND APPLICATION OF MASS SPECTROMETRY-BASED PROTEOMICS TO GENERATE AND NAVIGATE THE PROTEOMES OF THE GENUS POPULUS

Paul Edward Abraham
pabraham@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Paul Edward Abraham entitled "DEVELOPMENT AND APPLICATION OF MASS SPECTROMETRY-BASED PROTEOMICS TO GENERATE AND NAVIGATE THE PROTEOMES OF THE GENUS POPULUS." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Robert Hettich, Major Professor

We have read this dissertation and recommend its acceptance:

Mircea Podar, Steven Wilhelm, Albrecht von Arnim, Loren Hauser

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

DEVELOPMENT AND APPLICATION OF MASS SPECTROMETRY-BASED PROTEOMICS TO GENERATE AND NAVIGATE THE PROTEOMES OF THE GENUS *POPULUS*

A Dissertation Presented for the
Doctor of Philosophy Degree
The University of Tennessee, Knoxville

Paul Edward Abraham
May 2013

ACKNOWLEDGEMENTS

First, I'd like to gratefully and sincerely thank Dr. Robert Hettich for guiding me through the completion of my dissertation and subsequent Ph.D. Bob's open door policy, relaxed demeanor, and constant encouragement made for a good working relationship. His mentorship was paramount in providing a well-rounded graduate school experience. I am grateful for all of the opportunities he has provided, especially all of those "vacations" at the annual American Society for Mass Spectrometry conferences. Over these past 7 years at Oak Ridge National Laboratory (two of those spent as an undergraduate research assistant), I've learned so much from Bob and I could not thank him enough. If imitation is the sincerest form of flattery and gratitude, I'd like to highlight a few of his colorful ways of giving advice (a.k.a. Bobisms): team work ("We are all pegs on a stool", "We are all in the same boat and we shoot the swimmers", and "A rising tide lifts all boats"), life's complications ("A fly in the ointment", "Life is a complicated landscape", "It's all dynamic", and "It's a moving target", "Tipping your hand", "Chopping at the bit", and "Raise the red herring up the flagpole"), prioritization ("Too many irons in the fire, "Your dance card is full", "What's on the docket", and "We are going to build a white fence so that if you reach outside, you get your arm chopped off"), and a manuscript in progress ("Put all your eggs in a basket and shoot for the homerun fence", "Don't throw the baby out with the bathwater", "First show me the baby and then you can tell me about the labor pains", "The onus is on us", "An albatross around our necks", and "From soup to nuts").

Secondly, I'm very grateful to the remaining members of my dissertation committee, Dr. Albrecht von Arnim, Dr. Loren Hauser, Dr. Mircea Podar, and Dr. Steven Wilhelm. Their academic support and input are greatly appreciated. I'd also like to express my gratitude to Dr. Gerald Tuskan for his mentorship, guidance and support. My gratitude is also extended to other members of Bob's research group. Specifically, I'd like to thank Dr. Nathan Verberkmoes, Dr. Richard Giannone, Dr. Brian Erickson, Dr. Alison Erickson, Dr. Andrew Dykstra, and Dr. Rachel Adams for their guidance, discussions and suggestions throughout my graduate career.

Thirdly, I would like to thank my parents Paul and Cindy Abraham. They have both instilled many qualities in me, giving me a foundation with which to live honestly, responsibly, and compassionately. I would also like to thank my brothers Andy and Ryan Abraham, my little sister Leslie Abraham, as well as my mother-in-law Terry Lynn Hursey. I would also like to thank the three best friends that anyone could have, Jason Stewart, Brandon Barger, and Dr. Andy Harris. All of you helped shape me into the person I am today. Thank you.

Finally, and most importantly, my upmost gratitude goes to my beautiful wife Annie. Her encouragement, support, patience and constant selflessness were undeniably the foundation upon which the past four years of my life have been built. Although she endured many, many long hours alone while I worked on my dissertation, I was always greeted with a smile. Last, but certainly not least, I must acknowledge my son Parker Edward Abraham. At only 4 months old, you inspired me to never surrender the goals I have set.

ABSTRACT

Historically, there has been tremendous synergy between biology and analytical technology, such that one drives the development of the other. Over the past two decades, their interrelatedness has catalyzed entirely new experimental approaches and unlocked new types of biological questions, as exemplified by the advancements of the field of mass spectrometry (MS)-based proteomics. MS-based proteomics, which provides a more complete measurement of all the proteins in a cell, has revolutionized a variety of scientific fields, ranging from characterizing proteins expressed by a microorganism to tracking cancer-related biomarkers. Though MS technology has advanced significantly, the analysis of complicated proteomes, such as plants or humans, remains challenging because of the incongruity between the complexity of the biological samples and the analytical techniques available. In this dissertation, analytical methods utilizing state-of-the-art MS instrumentation have been developed to address challenges associated with both qualitative and quantitative characterization of eukaryotic organisms. In particular, these efforts focus on characterizing *Populus*, a model organism and potential feedstock for bioenergy. The effectiveness of pre-existing MS techniques, initially developed to identify proteins reliably in microbial proteomes, were tested to define the boundaries and characterize the landscape of functional genome expression in *Populus*. Although these approaches were generally successful, achieving maximal proteome coverage was still limited by a number of factors, including genome complexity, the dynamic range of protein identification, and the abundance of protein variants. To overcome these challenges, improvements were needed in sample preparation, MS instrumentation, and bioinformatics. Optimization of experimental procedures and implementation of current state-of-the-art instrumentation afforded the most detailed look into the predicted proteome space of *Populus*, offering varying proteome perspectives: 1) network-wide, 2) pathway-specific, and 3) protein-level viewpoints. In addition, we implemented two bioinformatic approaches that were capable of decoding the plasticity of the *Populus* proteome, facilitating the identification of single amino acid polymorphisms and generating a more accurate profile of protein expression. Though the methods and results presented in this dissertation have direct implications in the study of bioenergy research, more broadly this dissertation focuses on developing techniques to contend with the notorious challenges associated with protein characterization in all eukaryotic organisms.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

PRINCIPLES OF MASS SPECTROMETRY-BASED PROTEOMICS AND ITS APPLICATIONS TO A BIOENERGY CROP

1.1 The Omics Era: Enabling a Systems-level View of How Cells Function

A living cell can be described as a tightly regulated, yet readily adaptable system that contains a collection of molecules that work synergistically to dictate its function and organization, allowing biological processes such as cell growth and adaptation to varying environmental perturbations to occur simultaneously. Understanding the molecular basis of how cells function is a fundamental goal of molecular biology and is crucial to improving human health (i.e., aberrations in biological process play a role in diseases such as Parkinson's disease) and the environment (i.e., bioremediation strategies utilize key biological processes to reduce pollution). The birth of modern, molecular biology began in 1958; the discovery that each cell contains a molecule encoding all of the genetic information necessary for an organism to persist had profound implications for molecular biology and is considered one of the greatest scientific achievements of the twentieth century[1-2]. The genetic material identified was DNA, an abbreviation for deoxyribonucleic acid. This discovery led to the central dogma of molecular biology, which described a unidirectional flow of information from DNA to messenger ribonucleic acid (mRNA) to proteins[3]. Since the establishment of the central dogma, the increasing use of high-throughput analytical technologies has advanced our understanding of the relationship between these three molecules and concomitantly, heralded a new way of thinking about molecular biology – the "omics" era.

After the initial discovery of the double-helix structure of the DNA molecule, scientists could now examine the genetic code embedded within it; the most critical feature of a DNA molecule. Once the sequence of a DNA molecule is deciphered, scientists' could identify the genes that it contains and study their activities in greater detail. Beginning in 1960s, the coding race began and molecular biologists scrambled to

be the first to decipher life's code. By 1965, the language of DNA was understood and since then, DNA molecular biologists have been able to develop techniques to obtain longer stretches of DNA sequencing, culminating in the 1995 completion of the first whole-genome (i.e., knowledge of the complete set of coding and non-coding DNA) sequence of the bacterium *Haemophilus influenzae*[4]. Subsequently, the success of the first whole-genome sequencing led to the draft of the first human genome in 2001[5-6]. The information gleaned from a sequenced genome enabled scientists to explore an organism's genetic space, giving scientists' the ability to measure the genotype of any organism (i.e., the field of genomics)[7]. Making use of constantly refined technologies, the number of publicly available genome sequences is rising exponentially. Today, almost 2,000 genomes have been completed. Of those sequenced, 1644 are prokaryotic species or strains, 117 are archaeal and 153 are eukaryotic. In the present genomic era, as a testament to the rate of DNA sequencing, there are over 11,000 ongoing projects[8]. With the availability of high-quality genome catalogs, attention is now being directed towards the transcriptome (i.e. the complete set of transcripts in a cell) and proteome (i.e., the complete set of protein isoforms in a cell), which are a means to understanding how the information contained in the genome is used by the cell. This paradigm shift has given rise to *functional genomics*, which endeavors to make use of the massive amount of information to understand how genes actually function at the cellular-level[9-10].

Instead of focusing on the static physical aspects of the genome (i.e., the presences or absence of genetic content), functional genomics aims to understand the dynamic aspects of gene function at the level of transcripts and proteins, albeit with the genome as an anchor point. To understand the functional elements of the genome, the study of the transcriptome is a natural starting point, since genes are expressed through RNA transcription. When a cell enters a specific functional role, only a selective set of genes are transcribed into mRNA molecules (i.e., the RNA molecules that convey genetic information to the ribosome), and thus transcriptomics has emerged as a powerful approach for surveying gene expression. The primary aims of transcriptomic investigations are to catalogue all the transcripts during different cellular states in order to determine the transcriptional structure of genes and quantify the expression levels of each

transcript. Among the various technologies that have been developed to deduce and quantify the transcriptome, hybridization approaches (i.e., microarray-based transcription profiling) have been the mainstream technology for the past decade. When a genome sequence is available, genome tiling microarrays prove most effective in profiling gene expression because they offer a less biased survey of mRNA transcripts within cells[11-13]. Despite how successful this approach has been at elucidating and interrogating patterns of mRNA transcripts within cells, the advent of accessible next-generation sequencing technology promises to offer a far more precise measurement of transcriptomes. To date, several studies comparing hybridization arrays to RNA-sequencing have been performed[14]. In comparison to microarrays, RNA-sequencing has many clear advantages, the most important being: better dynamic range, low technical variation and the ability to detect novel transcripts[15]. Though RNA-sequencing technology promises to provide a better understanding of spatio-temporal transcription profiles, one thing, however, that a transcriptomic catalog will not answer is the exact concentration (i.e., the presence of a protein) or activity of proteins, the final product of gene expression.

As discussed above, genomic and transcriptomic investigations are a rich source of information; however, neither a static genome nor the presence of a transcript can be used to measure the actual functional state of a cell at a particular time point. It is the proteins, not the genes or transcripts, which are directly responsible for the observed phenotype (i.e., the morphology, anatomy, and function of a cell). Within a cell, proteins catalyze and essentially control all biological processes. Hence, the study of proteins is not only a necessary goal of molecular biology but essential to fully understand gene function at a holistic, systems-level perspective. First coined in 1996, *proteomics* has emerged as an indispensable level of information in the functional genomics era[16]. Despite superficial similarities with DNA and RNA molecules, initial efforts to globally characterize proteins were analytically challenging, in part owing to the diverse physicochemical properties of amino acids, which are the chemical building blocks of protein molecules. Like other emerging fields of research, the major impediment was technological. In the 1970s, proteomics advanced considerably because, for the first time, scientists' could simultaneously resolve thousands of proteins via two-dimensional gel

(2-DE) electrophoresis[17-18]. In the most common implementation, protein molecules are separated by charge using isoelectric focusing (first dimension) and then by size (second dimension) using SDS polyacrylamide gel electrophoresis. Next, separated proteins are detected by staining, resulting in a two-dimensional image of proteins occupying a specific x- and y- coordinate. For each "spot", the staining intensity provides an estimate of the quantity of the protein(s) present. With this technique, the spot profiles from different biological samples could be compared, providing a general method to profile gene expression. Although scientists could analyze various protein expression patterns, this methodology was essentially descriptive and did not reveal the identity of the resolved proteins. The following decade, however, witnessed an accelerated pace of technological developments on this front. By the early 1990s, protein sequencing techniques, mass spectrometry-based approaches in particular, had fundamentally changed protein characterization in molecular biology.

1.2 Mass Spectrometry-based Proteomics

Today, mass spectrometry (MS) plays a pivotal role in proteomics, primarily as a consequence of one technical breakthrough in the late 1980s. This breakthrough – two ionization methods: matrix assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) - solved the difficult problem of transferring relatively large, non-volatile molecules into the gas phase in an electrically charged form, a process known as ionization[19-20]. John Fenn, who shared the 2002 Nobel Prize for chemistry, captured the significance of this accomplishment in his famous phrase, "We learned how to make elephants fly". After these breakthroughs, the development of low-cost commercial mass spectrometers equipped with either a MALDI or ESI ion source quickly followed, giving molecular biologists the ability to robustly measure and identify proteins. With proteins easily transferred into the gas phase, scientists could then measure the mass or, more precisely, the mass-to-charge ratio (*m/z*) by accelerating the ions to a mass analyzer. In general, MALDI ion sources were coupled with time-of-flight (TOF) mass analyzers[21], whereas ESI was commonly coupled with a triple-quadrupole[22] or ion-trap mass spectrometers[23]. While both methods allowed routine analysis of proteins, the

two techniques had divergent - but equally significant – reasons for their popularity. MALDI-TOF instruments rapidly gained popularity for a variety of reasons, mainly owing to the ease of use (i.e., the platform predominately generates singly-charged species, hence the results are relatively easy to interpret), broad mass range (i.e., m/z values could range from 400 to 350,000), high mass accuracy, and measurement sensitivity (i.e., capable of detecting low molecular quantities in a sample)[24]. With these figures of merit, the MALDI-TOF platform became a powerful tool to determine the molecular weight of intact proteins (i.e. top-down mass spectrometry) with a high degree of accuracy[25]. As a complement to MALDI-MS platform, ESI gained immediate popularity because the ionization process had the propensity to generate multiply-charged ions. In contrast to singly-charged ions, multiply-charged ions respond well to fragmentation processes, making the ESI ion source well-suited for use in tandem mass spectrometry (MS/MS) applications (tandem-in-time, ion trap; tandem-in-space, triple quadrupole) such as peptide sequencing[26]. Perhaps an even more important reason for ESI popularity was because the ease at which the source could interface liquid phase chromatography to mass spectrometers. Together, these features clearly established ESI-LC-MS/MS as a powerful platform for the analysis and subsequent identification of peptides in complex mixtures (bottom-up mass spectrometry). Since the establishment of these two complementary - and equally compelling – approaches, top-down and bottom-up mass spectrometry have evolved significantly. As these methods improved and were combined with genome sequencing, it was quickly recognized that these protein measurements could be used to comprehensively characterize proteomes at a larger scale.

Over the past two decades, top-down (TD) proteomics has earned a remarkable place in proteomics – driven by the interest in interrogating protein structure and identifying protein heterogeneity (i.e., the ability to study post-translational modifications, splice variants, mutations, etc.)[27]. In the years since its emergence, more rigorous experiments are being conducted. Rather than only measuring the intact molecular weight of a protein, direct fragmentation of the protein in the gas phase is routinely performed for partial sequencing. The first demonstration of the top-down fragmentation approach occurred in the 1990s, when a group of researchers measured the

electrosprayed ions of ribonuclease A using a triple quadrupole instrument[28]. In the years that followed, improvements in mass analyzer technology, specifically the introduction of Fourier transform ion cyclotron resonance (FTCIR), led to the first report of a protein ion analysis with high resolution (i.e., isotopic resolution)[29-30]. Today, the evolution of FT and TOF instruments have made high-resolution MS broadly available – a mass resolving power (i.e., full width at half height) of 100,000 is readily achievable with commercial instrumentation[31-32]. Emphasizing work with protein profiling, the advancements in MS hardware continue to improve TD LC-MS/MS approaches – the identification of 1,000-3000 proteins from complex mixtures is now possible[33]. To date, high-throughput TD proteomics is feasible: it has been applied for both discovery-focused as well as hypothesis-driven investigations and has been applied to various cell-types, ranging from microorganisms to cancer lines[34-35]. Although TD has advanced significantly, many challenges still remain. Currently, the greatest hindrance in TD proteomics is the difficulty in analyzing proteins with high molecular weights (>60 kDa)[36]. In addition to technological limitations, the front-end separation of intact proteins is more challenging than the separation of peptide mixtures analyzed by the bottom-up approach[37]. Moreover, there is a severe deficiency in algorithms and software that interpret results, especially compared to the variety of informatics tools available for bottom-up proteomic analysis. This discrepancy between top-down and bottom-up proteomics is largely in response to the more rapid advancement of techniques and instrumentation of bottom-up mass spectrometry, furthering its widespread adoption and development.

As discussed above, TD proteomics is a relatively primitive field compared to bottom-up proteomics. Given the complementary nature of the information provided by TD strategies, the approach will continue to be employed. However, until improvements in MS hardware and bioinformatic tools are made, the bottom-up (BU), or "shotgun", approach will remain the mainstream method when tackling high-complexity samples for large-scale proteomic investigations[38]. The term "shotgun proteomics" is frequently used to describe the method because it's analogous to shotgun genomic sequencing - in which the DNA is shredded into smaller parts, sequenced individually, and reassembled into their original order. Most BU proteomic applications rely on the proteolytic digestion of

proteins into peptides prior to mass analysis, followed by their subsequent introduction into a mass spectrometer. In the early 1990s, the first BU techniques were being routinely used to produce accurate peptide molecular weight "fingerprints"[39]. In conjunction with computational tools, this method identified proteins by matching the measured masses of two or more peptide fingerprints to theoretical peptide masses generated from a protein sequence database (i.e., information obtained from genome sequencing). For simple protein mixtures, like those obtained from 2-DE, accurate peptide mass information provided enough discrimination to identify unknown proteins "on the fly". In 1993, peptide mass fingerprinting was coupled with liquid chromatography (LC) to further separate peptide mixtures - this significantly improved the overall number of peptides observed and marked the beginning of the proteomics era[40]. However, when dealing with more heterogeneous peptide mixtures, peptide mass fingerprints alone were generally not accepted as sufficient evidence for protein identification – peptides may have the same mass.

Around the same time, Mann and Yates demonstrated a new approach to connect mass spectrometry data with protein sequence databases[41-42]. Using tandem mass (MS/MS) spectrometry, peptide molecular masses, along with their amino acid sequences, are obtained and, to this day, have proven more useful for protein identification. In brief, MS/MS analysis is a two-step process. The first step involves recording the *m/z* values of all the peptides ions that are introduced into the mass analyzer at a given time (MS spectrum). Next, a single selected peptide *m/z* (often referred to as the "precursor" ion) is fragmented into smaller pieces (fragment ions) in the collision cell of the mass spectrometer. A MS/MS spectrum is therefore a record of all the fragment ion *m/z* values generated from an isolated precursor ion. In practice, MS/MS spectra are deciphered by identifying a consecutive series of fragment ions whose differences correspond to molecular masses of amino acids. As such, the fragmentation pattern encoded by a MS/MS spectrum allows the identification of the peptide that produced it. Using this method, Mann and coworkers demonstrated that, even though the interpretation of MS/MS spectra is complex, one could easily identify "runs" of fragment ions, which yield partial sequences called "peptide sequence tags"[41]. By using peptide

sequences tags, peptides could then be located to a specific protein in a sequence database. The introduction of the database search algorithm, PeptideSearch, allowed for rapid error-tolerant identification of peptides in protein sequence databases. A parallel development by John Yates and coworkers searched MS/MS spectra with a different approach, based on cross-correlation of a theoretical spectrum with the experimental fragment ion spectrum, to identify proteins[42]. Since peptides fragment in a highly predictable way, a search algorithm, named SEQUEST, was built with a cross-correlation function that provided a measurement of similarity between the fragment ions observed in a MS/MS spectrum and the predicted fragmentation patterns of peptide sequences in a database. Together, the Mann and Yates database search algorithms provided the necessary bioinformatic infrastructure to automatically identify peptides from MS/MS spectra. Two decades later, these approaches remain the basis for most, if not all, discovery-based BU investigations.

1.3 Bottom-up Proteomics for Discovery-based Investigations

1.3.1 Peptide Sequencing by Tandem Mass Spectrometry

During the 1970s and 1980s, key developments in instrumentation and progress in understanding gas-phase chemistry allowed for the mass analysis of precursor ions, fragmentation of the selected ions, and mass analysis of the resulting fragmentation ions (i.e. tandem mass spectrometry)[43]. During this period, Donald Hunt and colleagues were among the first to develop a peptide sequencing strategy based on tandem mass spectrometry[26]. Using a triple quadrupole mass spectrometer, Hunt showed how to derive peptide sequence information by fragmenting precursor ions with low energy (<200 eV) collision induced dissociation (CID)[44]. With this strategy, the ability to precisely select and sequence co-ionizing peptides became available. Soon thereafter, this fragmentation process was commercialized into the tandem-in-time instruments (ion selection and its dissociation occurs within the same space) such as the ion trap mass spectrometers, particularly the LCQ by the Finnigan Corporation. Compared to the triple quadrupole instrument, ion trap instruments provided much better precursor selection resolution and product ion resolution[45]. In the years to follow, tandem mass spectrometry via CID

fragmentation was, and still is, used ubiquitously for peptide sequencing, and thus will be discussed in greater detail.

As discussed previously, the MS/MS data acquisition process consists of two stages: (1) the instrument scans all peptide ions that are introduced into the mass analyzer at any given time and records the full-scan (or MS^1) spectrum – a list of the m/z ratios and intensities of all peptide ions; (2) a peptide ion observed in the MS1 spectrum is then isolated and fragmented by CID to break down the peptide into smaller pieces (Figure 1.1A-C). The acquired MS/MS (or MS^2) spectrum is thus a record of the *m/z* values and intensities of each fragment ion. The CID fragmentation pattern encoded by the MS/MS spectrum allows the identification of the amino acid sequence of the peptide that produced it. However, identifying a peptide sequence in an MS/MS spectrum is analogous to solving a jigsaw puzzle. The solution to the puzzle demands knowledge of how the peptides fragment and hence requires an understanding of the CID peptide fragmentation process. Fundamentally, CID fragmentation can be described as a two-stage phenomenon. In the first stage, fast-moving ions are activated (excited) through multiple energetic collisions with a neutral gas atom, such as helium or argon. Upon collision, a fraction of the ion's kinetic energy is transferred into internal vibrational energy. During the second stage, the rapidly deposited internal energy - caused by multiple collisions - subsequently promotes dissociation of the molecule into smaller fragments. Although the energy is randomly distributed across all atomic bonds, it is the weakest bond that breaks. For peptide ions, the dissociation pathways observed in CID spectra have been rationalized by the "mobile proton" model[46-47]. Since its establishment, the mobile proton model has provided a qualitative framework that permits the interpretation of peptide MS/MS spectra. Essentially, the model assumes that when protonated peptides are created by ionization methods such as ESI, the protons are initially localized on the most basic sites of the ion. In a peptide, these sites are the N-terminus and the side chains of basic amino acid residues (i.e., histidine, lysine, and arginine).

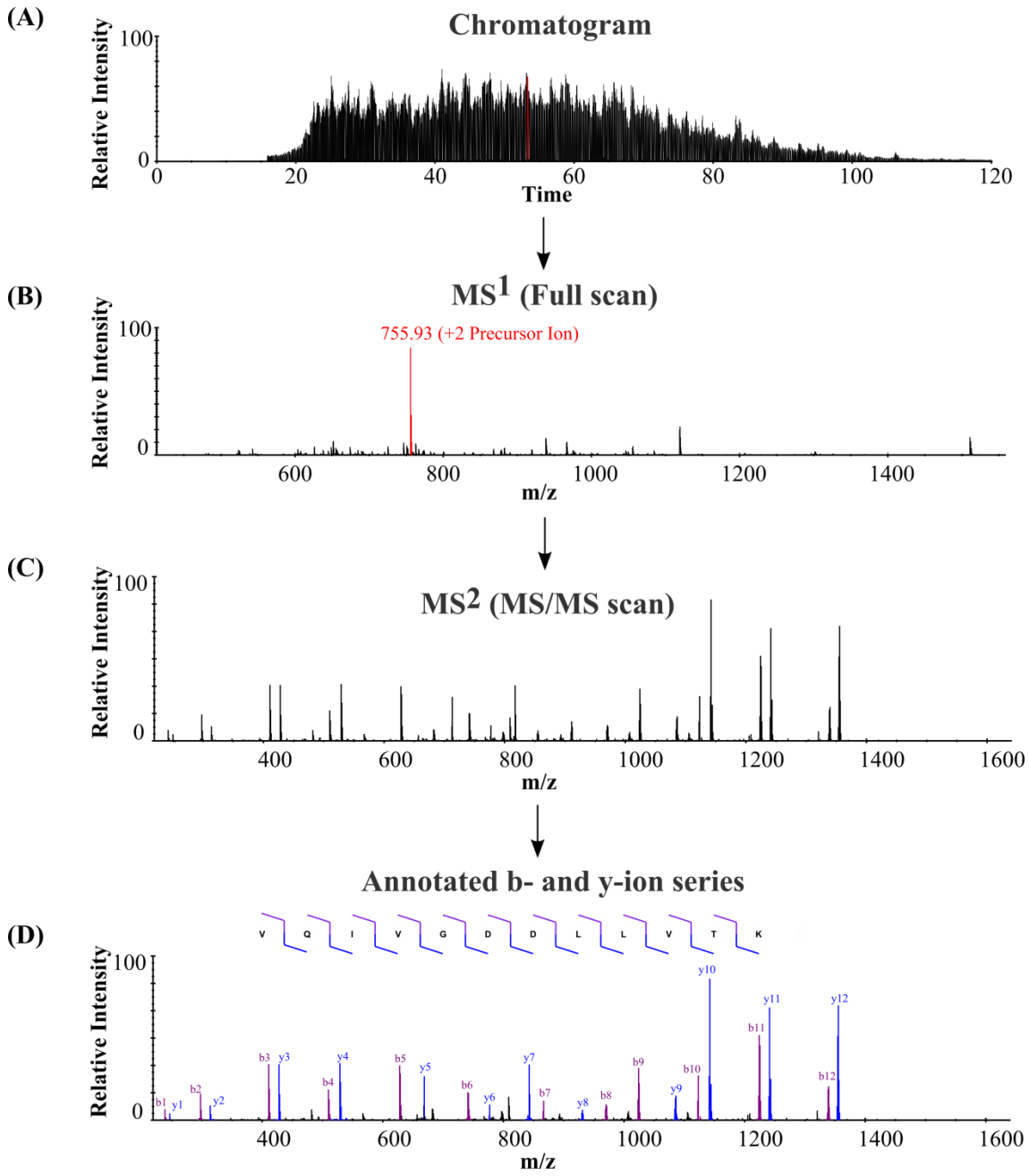


Figure 1.1. Illustration of how peptides sequencing information is obtained during a tandem mass spectrometry experiment. (A) Peptides are separated by liquid chromatography and the *m/z* values are observed in a (B) MS^1 spectrum. A single precursor ion is isolated for CID fragmentation, which produces a (C) MS^2 spectrum. (D) The spectral information in each MS2 spectrum can be used to identify the peptide sequence. For CID, the b- and y-ion series is used.

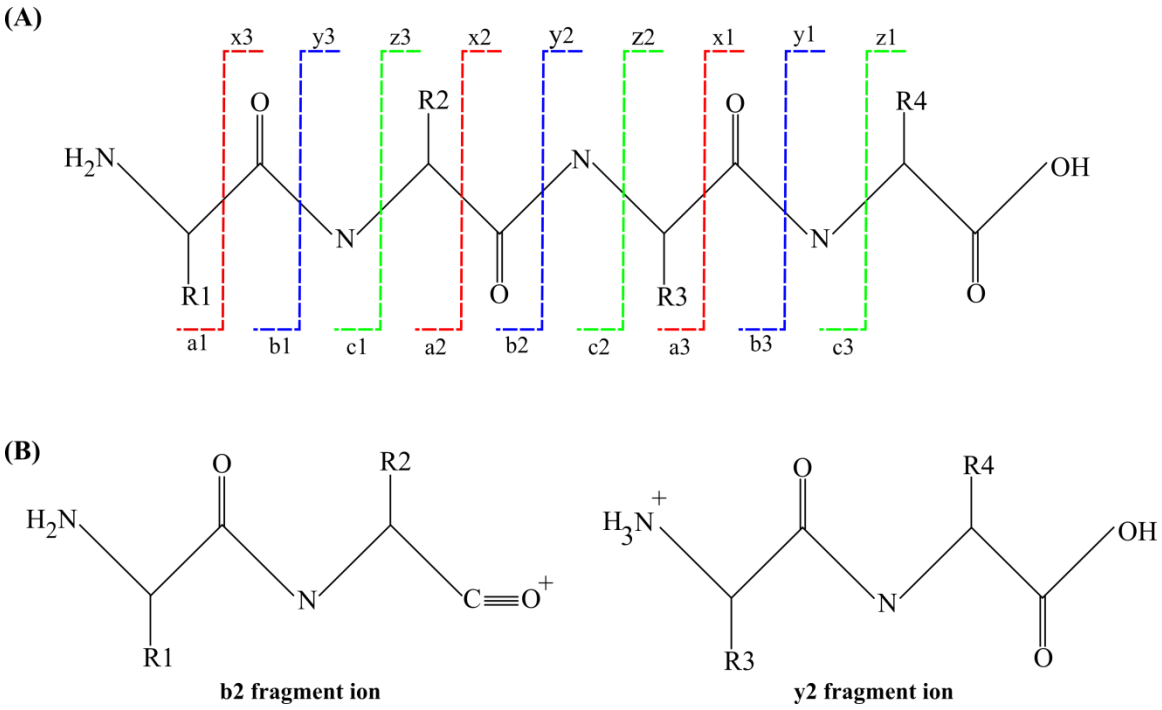


Figure 1.2. Illustration of collision induced fragmentation of a polypeptide. (A) A peptide backbone with four amino acid residues (R) and the types of fragment ions generated in a CID MS/MS spectrum. If the charge is retained on the N-terminal side, the fragment ion is classified as either a, b, or c. If the charge is retained on the C-terminal side, the ion type is x, y, or z. (B) The fragment ion structure of a b- and y-type fragment ion.

After the peptide ion becomes activated in a collision cell, the additional energy allows the proton(s) to explore less-basic sites along the peptide chain. Among the potential locations that can be broken by these collisions, protonation most often occurs at the amide site (Figure 1.2)[48]. Protonation of the neutral amide bond weakens its stability, thus permitting the cleavage of the peptide bond and creating b- and y-ions, which denotes charge retention on either the N- or C-terminus. Since CID fragmentation occurs more or less randomly at various amide bonds across a peptide backbone, a series ("ladder") of b- and y-ions will be present in each MS/MS spectrum. This fragmentation ladder, in fact, can be manually interpreted to identify a peptide amino acid sequence, albeit with efficient training in the interpretation of MS/MS spectra. Manual analysis is done by looking at the mass difference between peaks of a MS/MS spectrum and determining if the mass difference corresponds to an amino acid residue – identifying consecutive ions belonging to the same series (b- or y- ion series) allows the determination of the peptide's sequence (Figure 1.1D). In principle, complete coverage of either b- or y- type ions allows full annotation of the entire peptide amino acid sequence from a MS/MS spectrum. However, complete coverage of tandem mass spectra – particularly for larger peptides - is rarely achieved, resulting in missing fragment ions, which produces gaps in the analyzed amino acid sequence. This result can cause an incorrect identification, especially if only relatively few peaks are annotated. Due to the scale of modern proteomic experiments (over 100,000 MS/MS spectra are typically collected), manual analysis of collected MS/MS spectra is clearly impractical. Given this limitation, computer algorithms have been constructed - based on expert knowledge of CID fragmentation – to annotate MS/MS spectra in an automated fashion (*vide infra*).

1.3.2 Peptide Identification by Searching Algorithms

The rapid identification of peptides, using tandem mass spectra and sequencing algorithms, ushered in a new era of bottom-up proteomics. By simplifying the most time-consuming part (i.e., annotation of MS/MS spectra), computational approaches that automatically assigned peptide sequences to MS/MS spectra became an essential tool for large-scale protein analyses. Not only did these algorithms increase throughput, the ability to automate the interpretation of MS/MS spectra opened the door for non-experts

to perform the analyses. Today, nearly all tandem mass spectra are interpreted by three types of algorithms: 1) *de novo* sequencing, where the amino acid sequence of a peptide is explicitly "read out" from a MS/MS spectrum; 2) a hybrid approach that combines *de novo* sequencing and database searching, where short peptide sequence tags of 3-5 residues in length are extracted from MS/MS spectra and used for error-tolerant database searching; and 3) database searching, where peptides are identified by correlating acquired experimental MS/MS spectra with theoretical spectra predicted for each peptide contained in a protein sequence database.

Due to the lack of genomic sequencing information, *de novo* analyses preceded the other sequencing strategies by several decades - the concept extends back to experiments performed by Klaus Biemann in the early 1970s[49]. Over the course of subsequent decades, numerous *de novo* sequencing algorithms have been developed to assist researchers in this task. Among these programs, the better known are PEAKS[50], MSNovo[51], and PepNovo[52], which use sophisticated probabilistic scoring functions that measure the matching quality between a MS/MS spectrum and a peptide sequence. The primary advantage of the *de novo* approach for peptide sequencing is that each spectrum is given an equal opportunity to match any combination of amino acid residues, regardless of whether the researcher anticipated detecting the sequence or not. The independence to a predicted sequence database makes *de novo* sequencing the preferred method to identify unexpected peptide modifications (for example, peptides containing post-translational modifications, truncations, or mutations) and study species for which no or limited genome sequence information is available. Despite this attractive capability, *de novo* sequencing has not yet become a practically useful approach for large-scale data analysis. In fact, what makes *de novo* sequencing so attractive also makes it extremely challenging. Due to the lack of a reference database, the analysis requires MS/MS spectra of the highest quality and more advanced algorithms, which are computationally intensive[53]. Moreover, *de novo* sequencing greatly increases the number of candidate peptide sequences compared to each spectrum, consequently incurring not only significant costs to processing time but also unacceptable error rates. With the advent of high-through genome sequencing, this approached has adapted appropriately. Today, a

number of approaches combine *de novo* sequencing elements with database searching algorithms.

First pioneered by Matthias Mann in 1994[41], the analysis starts with the extraction of short sequence tags (partial sequences) from each MS/MS spectrum. An error-tolerant database search for each MS/MS spectrum is then performed, importantly, only against database peptides that contain one of the partial sequences, thereby reducing the processing time and error rates. Moreover, these searches can be augmented to allow for one or more mismatches between the sequence inferred from the MS/MS spectrum and the database sequence. This powerful concept has been extended by several groups to systematically identify chemically modified peptides (i.e., post-translational modifications) or peptides containing mutations (i.e., cancer biomarkers)[54-57]. By uniting *de novo sequencing* with database tag reconciliations, this hybrid approach remains extremely relevant in modern proteomics-based research.

Today, database searching algorithms remain the most frequently used method for large-scale protein identification. Since 1994, numerous commercial and open-source database search algorithms have been developed[42, 58-64]. From a simplified view, these algorithms perform the same basic functions - these programs take MS/MS spectra as input and score them against theoretical fragmentation patterns constructed from peptides in a database. For each query spectrum, the search algorithm will apply user-specified criteria (for example, mass error tolerance) to restrict the matching process to a confined list of candidate peptide sequences. Next, a set of expected fragment ions are calculated for each candidate peptide sequence and then compared against the experimental spectrum. A search score is then calculated for each peptide-spectrum match (i.e., a measure of the degree of similarity between the experimental MS/MS spectrum and the theoretical spectrum). There are a number of scoring schemes that have been described and implemented into popular searching algorithms, including correlation functions (SEQUEST[42], MASCOT[58], MyriMatch[63], and OMSSA[64]), dot products (SpectraST[65]), and probabilistic fragmentation frequencies (PHENYX[62]). Although empirical evidence does suggest certain search algorithms outperform others, the overlap between different algorithms is often in the range of 70%[66]. These search scores, nevertheless, serve as a

primary discriminating parameter for separating correct from incorrect identifications. In certain proteome data sets, especially those generating low quality spectra (i.e., low mass accuracy instrumentation), a search algorithm can produce a peptide match for almost every MS/MS spectrum. Therefore, it is important to stress that a search score alone cannot be used as a reliable indicator for a true match. In order to discriminate true from false matches, methods that statistically assess peptide assignments have become a necessity.

Early on, spectral matching outputs would be refined to a list of identifications using *ad hoc* decision making, often in combination with manual inspection of the peptide assignments. This type of approach, however, makes comparisons between data sets essentially impossible and more importantly, visual inspection of MS/MS spectra is not a viable validation process because it's subjective and time-consuming. Therefore, modern proteomic studies use a more practical way to control erroneous peptide spectrum matches. Today, statistical approaches are used to provide a global assessment of confidences and estimated false discovery rates (FDRs). Pioneered by Hochberg and Benjamini[67], a false discovery rate is defined as the expected proportion of incorrect identifications among all identifications in the data set. In MS/MS-based proteomics, the most commonly used and accepted method for computing the FDR is the target-decoy strategy[68]. Simply, the strategy requires that all experimental MS/MS spectra are searched against the reference protein database (target) appended with a reversed, shuffled, or randomized version of itself (decoy). Here, the underlying assumption is that false matches to sequences from the target database will follow the same search score distribution as matches to decoy peptide sequences. Therefore, an entire data set can be filtered at various search score cut-offs, and a corresponding FDR can be computed as $2N_d/N_t$, where N_d is the number of decoy peptide matches and N_t is the total number of matches. Once an acceptable FDR is achieved and the peptide-spectrum matches have been statistically validated, protein identifications can be inferred, and perhaps quantified, from the peptides detected.

1.3.3 Protein Inference

Although bottom-up proteomics is a peptide-centric approach, the explicit goal is not the identification of peptides, but rather the identification of the proteins expressed in a cell. Therefore, every peptide sequence identified needs to be assigned to its corresponding protein. At present, a diverse set of computational approaches for assembling peptides into proteins have been developed[69-71]. In general, the process consists of the following steps: 1) protein sequences corresponding to each peptide are retrieved from the supplied database; 2) peptides are grouped by their corresponding protein sequences; 3) shared peptides are apportioned among all corresponding proteins. The plausibility of the reassembling process, however, is difficult to quantify, owing in part to what made bottom-up proteomics so successful (i.e., proteins were digested into peptides because they were easier to separate, ionize, and sequence). With the connectivity between peptides and proteins lost, numerous methods and philosophies have been proposed to define the criteria needed for calling a protein "identified" and still, there is no generally accepted way to do it.

Typically, the identification of a single peptide is sufficient evidence to identify a protein; however, it is often not enough to discriminate between two proteins that share extensive sequence homology, as is the case in higher eukaryotes such as plants and humans[72]. Therefore, the most accepted guideline among researchers requires at least one distinct peptide sequences (i.e., a peptide amino acid sequence that corresponds to one protein and no others in the database) for a protein to be identified. Currently, inconsistencies in protein assembly exist because there is no generally accepted way to handle shared peptides: peptides whose sequences are present in more than one protein[73]. For example, some research groups neglect this "protein inference problem" and just apportion shared peptide sequences to their corresponding proteins. With this approach, different proteins in a database could be counted as separate protein identifications even if they share the same set of peptides. In this scenario, these redundancies over-estimate the number of proteins present and, more importantly, lead to an incorrect interpretation of the data. To avoid this ambiguity all together, some research groups, on the other hand, disregard the protein inference problem entirely by eliminating shared peptides from the

data set. For microbial studies, this approach may not incur any significant loss in data since prokaryotic organisms have very few, if any, shared peptides. However, this type of approach can significantly under-estimate the protein content of highly redundant eukaryotic data sets (i.e., shared peptides resulting from whole-genome duplications, protein families or alternative splicing variants comprise a large fraction of total peptide library)[74]. Therefore, it has become increasingly recognized that there needs to be a nomenclature that provides a consistent way for presenting the results from large-scale proteomic investigations.

Today, the most commonly adopted nomenclature for protein classification couples the principles of parsimony with Occam's razor constraint. First described by Yang and colleagues[71], the approach rationally organizes the results into a minimal list of proteins that sufficiently explain all of the identified peptides. The nomenclature described classified proteins by their level of ambiguity: proteins that consist of only distinct peptides are classified as *distinct* proteins; proteins are classified as *differentiable* when they contain at least one peptide that is unique to that protein, as well as one or more peptides that map elsewhere in the proteome; *indistinguishable* proteins consist of only shared peptides. Since indistinguishable proteins are difficult to rationalize and interpret, most groups eliminate these from the data set. Presenting the results in terms of such minimal lists has several advantages: allows a consistent calculation of the number of proteins identified in the experiment and simplifies the interpretation to only those proteins that are conclusively determined to be present in the sample.

With a conclusive list of proteins present in a complex sample, this information can be used to assemble proteome maps: that is, an inventory of proteins identified for a particular proteome. Over the past decade, progress in achieving maximal proteome coverage has been remarkable (for example, 88% proteome coverage has been achieved for *M. mobile*[75], 81% for *M. pneumoniae*[76], 54% for *C. elegans*[77], 97% for *S. cerevisiae*[78], and 48% for *A. thaliana*[79]). Ten years ago, a study that could identify a few hundred proteins was considered a monumental achievement. Today, a typical bottom-up experiment can readily allow the identification of thousands of proteins. For organisms, like bacteria - which have smaller genomes - it is feasible to identify almost 70% of the

proteome in a single experiment. Despite recent advancements, complete coverage of a proteome has not been achieved. However, this is not surprising. Generating complete proteome maps is a difficult task as it is unlikely the entire ensemble of proteins encoded by a genome will be expressed at any given time. In a cell, gene and protein expression is state-specific and so complete proteome coverage would require a compilation of partial proteome maps from many different cellular states. For moderately complex organisms (i.e., plants and humans), this task becomes remains challenging, slow, and expensive. Although the discovery of proteins has led to a better understanding of key biological process, current efforts are geared toward understanding a protein's context in a biological system, for example measuring protein abundance as a cell responds to a changing environment.

1.3.4 Protein Quantitation

Compared to the complementary and more mature technologies of transcriptomics, MS-based proteomics is of central importance to the emerging systems-level biology era. Increasingly recognized, the measurement of differential protein expression has become a necessity because it provides a more direct, more accurate way to measure global changes in cellular dynamics[80]. Therefore, quantitative proteomics holds significant promise for the discovery of prognostic or diagnostic cancer biomarkers, therapeutics studies (i.e., proteomic changes observed when a cell state is perturbed by a particular drug) and as a powerful platform to further our understanding of key biological processes. Quantitative proteomics can provide either a comprehensive measure of relative protein abundances (fold-change) between two or more conditions or a more targeted measure of the absolute quantity of a protein (average number of protein copies per cell)[81]. To date, the two primary ways to quantitate proteins apply either a label-free or stable isotope-based method.

Historically, bottom-up proteomics has focused on maximizing the number of identified peptides and proteins. Apart from the qualitative information gained (i.e., precursor ion sequences), these discovery-based approaches simultaneously acquire quantitative information (i.e., relative precursor ion intensities). During data acquisition, peptide fragmentation events are often triggered by ion abundance levels. Consequently,

the number of times a peptide can be observed and subsequently sequenced is dependent on the width (retention time) and height (ion intensity) of its chromatographic peak. Since a peptide's ESI signal has been reported to correlate linearly with an increasing concentration of its protein, either the peak area[82-83] or the number of observations (i.e., spectral count)[84-86] can be used for relative quantitation of proteins across various samples. Over the past decade, label-free approaches have been applied to comprehensively measure relative changes of peptide - and indirectly the associated proteins – for different samples. It should be stressed that, since relative changes in peptide signals are used for quantitative comparisons, reproducibility of the MS measurements is critically important. In addition, the different biological samples under investigation must be processed and analyzed under rigorously controlled conditions. Finally, significant normalization procedures are required to reduce the effects of experimental and technical biases. Today, label-free strategies are the ideal quantitative method for robust, yet comprehensive comparative analyses.

For differential analysis between samples, another widely used quantitative approach, albeit more expensive and time-consuming, incorporates heavy versions of specific molecules into peptides, either by metabolic labeling or chemical derivatization. First introduced in 1998 by Langen and colleagues, metabolic labeling experiments "feed" cells a specific isotope media[87]. For microbes and plant species, the most common approach uses heavy nitrogen isotope (^{15}N) for labeling. For mammals, SILAC[88] is the method of choice – cells are "fed" heavy amino acids, for example a heavy amino acid can contain ^{13}C instead of ^{12}C. Both approaches are conceptually and experimentally similar: one sample contains heavy isotope-substituted amino acids, while another sample contains normal (light) amino acid residues. The labeled and unlabeled cells are mixed and the proteomes are extracted and measured by MS-based approaches. Since the incorporation of heavy isotopes into a peptide leads to an expected mass shift, each peptide appears as a pair in the full scan spectra: the peptide with the lower *m/z* value contains the lighter amino acids and the peptide with the higher *m/z* values contains the heavier amino acids. As a result, the ratio of the peptide ion intensities in the full scan spectrum directly yields the ratio of the proteins between cell populations. An advantage

of metabolic labeling is that the heavy and light cells are mixed prior to sample processing, eliminating any quantitative biases due to processing errors. A limitation of this approach is that it may not be amenable to cells that are difficult to grow or show sensitivity to media composition, as this may negatively impact the proteome. Moreover, another major disadvantage of these MS^1-based quantification methods is that the complexity of the MS^1 scan has been doubled, thereby reducing the sensitivity and the data acquisition of the analysis. Whereas metabolic labeling requires living cells, chemical labeling can be done on any proteome. For example, clinical samples (i.e., tissues or biological fluids) require chemical labeling methods. In these approaches, isotope-bearing "tags" are targeted toward reactive sites on a peptide or protein. Today, popular implementation of this approach targets cysteine residues (i.e., ICAT[89]) and amino groups (i.e., iTRAQ[90] and TMT[91]). In these methods, the determination of relative peptide abundances is performed on reporter fragment ions measured in the MS/MS spectrum. Importantly, this type of approach allows up to eighteen samples to be compared and measured concurrently (i.e., multiplexing), thereby increasing throughput potential[92].

While these comprehensive approaches are good at measuring large changes in protein expression, they are less reliable for small changes. In fact, there are experimental trade-offs when performing comprehensiveness differential analyses versus precise quantitative measurements - because depth is the primary interest of differential proteomic investigations, the overall analytical precision is reduced in terms of the limit of detection and dynamic range[93]. Moreover, quantification is significantly compromised in complex samples, for example ionization suppression impinges on the precision of the measurement. As such, these approaches are often referred to as semi-quantitative approaches – that is, the precise determination of the concentration of peptides in the sample has a coefficient of variation that's typically greater than 20%. For the absolute quantitation (AQUA) of proteins, quantitation is achieved via a targeted approach with any of the commonly used stable isotope labeling methods[94-95]. Although this strategy requires prior information of the targeted proteins through a discovery-based approach, this strategy, a targeted analysis affords higher selectivity, a lower limit of detection, and

a wider analytical dynamic range[80]. In contrast to relative quantitation, the approach is considerably more reproducible and high throughput (i.e., after the SRM assay conditions and parameters have been established). The typical implementation of the approach is single reaction monitoring (SRM) – also called multiple reaction monitoring (MRM) - strategies[96]. SRM exploits the unique capabilities of the triple quadrupole instrument to selectively record peptide fragmentation events over time for the precise quantification of a predefined and specific target peptide[97]. In an SRM experiment, the most critical step is the careful selection of target peptides for each protein of interest. For proper selection, the following criteria are of critical importance: the peptides must be MS compatible; it is essential that each peptide is unique to their respective protein; a peptide cannot be chemically modified (i.e., post-translational modifications); a peptide cannot contain a missed cleavage site. After selection, peptides are quantified using specific MS settings. For highly sensitive and selective quantitation, an SRM measurement filters peptides and their respective fragment ions using the first and third quadrupole. Here, the first quadrupole acts as a mass filter to specifically select the *m/z* values corresponding to a single peptide ion. The selected peptide ion then passes through the second quadrupole, which serves as a collision cell. Following CID fragmentation, the third quadrupole filters the *m/z* values of specific, predetermined fragment ions of the peptide. During the course of an experiment, several transitions (peptide ion/fragment ion pairs) are monitored over time, providing a record of the peptide's signal intensity with retention time (i.e., peak area). Together, these coordinates yield a definitive quantitative assay for the targeted peptide. Since the final goal of an SRM experiment is the precise quantification of a set of target proteins, the peak areas of at least three peptides are acquired to quantify a single protein. For absolute quantitation, a known concentration of an isotopically labeled (heavy) peptide is introduced into the sample and measured concomitantly with the lighter peptide, allowing a direct comparison of the peak areas of the isotopically labeled peptide and the lighter peptide. In such an approach, one can readily derive accurate protein concentrations – even for proteins that are less than 50 copies per cell - with coefficient of variations less than 20%[98]. With this capability, targeted approaches will

remain a key technology for hypothesis driven investigations aimed at determining the concentrations of specific proteins[99].

1.4 Proteomics of a Bioenergy-relevant Organism

1.4.1 Introduction to Bioenergy

In recent years, increasing economic and environmental concerns - associated with the dependency on fossil fuels - has led to the introduction of government policies supporting bioenergy research[100-101]. Generally speaking, bioenergy research aims to find ways to extract sustainable and renewable energy (i.e., hydrogen, fuels, or electricity) from biological sources (i.e., biomass). Today, enormous efforts are being performed to improve the biological conversion of plant biomass to a sustainable energy supply – plant biomass is converted to biomass by breaking down the cell wall carbohydrates to simple sugars, which can be fermented to ethanol[102]. In particular, many aim to improve the production of bioethanol from a variety of feedstocks (for example, agricultural and forestry). At present, bioenergy is making a substantial contribution to reducing our dependency on gasoline – conversion of conventional starch crops, such as corn, to ethanol is well established and commercialized[103]. Although sugar (i.e., glucose) can be readily extracted from corn and converted to ethanol with existing technology, the current chemical energy yield is low: about one-third of the chemical energy is lost in producing ethanol[104]. Moreover, corn-based ethanol production has raised many environmental and economic concerns, such as fertilizer runoff, potent greenhouse gases, and fuel competing with food. While such issues can be mitigated by regulations and technical advancements, a major effort has begun to develop alternative feedstocks for ethanol by using crop residues, perennial grasses, and other forms of plant biomass that are collectively termed "lignocellulosics"[105].

The use of these non-food bioenergy feedstocks would significantly decrease the potential social and economic issues associated with land use. If fact, several lignocellulosic crops can be grown on less favorable soils and climate conditions. When considering the net energy output of these feedstocks, the bioenergy output per hectare is much larger than conventional crops for a number of reasons: lignocellulosic crops are

often perennials, they require fewer agronomic inputs, and a higher percentage of the harvested biomass can be used for ethanol production[106]. While the favorable features of lignocellulosic feedstocks offers prospects of low costs, improved efficiencies, and improved greenhouse gas emissions, current bioenergy efforts cannot take advantage of these feedstocks because they pose a unique set of challenges. Most importantly, this form of biomass presents a challenge in the accessibility to its simple sugars – lignocellulosic biomass cannot be readily converted to ethanol because the simple sugars are locked in a complex polymer composite that consists of a mixture of lignin, hemi-cellulose and cellulosic fibers[107]. As such, the current cost involved in cellulosic ethanol production is not competitive with the cost of oil[108-109]. By solving the recalcitrance of lignocellulose, there will be a dramatic cut in costs, which will in turn allow the renewable source of energy to be implemented into current fuel infrastructure. To make bioenergy processing viable, research efforts are currently dedicated to understanding how plant cell-wall molecular and physical structures are synthesized, maintained, and deconstructed. In almost every one of these processes it is the enzymatic catalysis, molecular signaling, and physical interactions of proteins that reflect the biological and chemical activity of the cell under various conditions. By identifying the proteins responsible for plant cell composition - particularly those that influence carbon allocation and carbon partitioning – within and among plant cells, variations in lignin composition can then be, in theory, controlled through genetic manipulations, thereby improving the conversion of lignocellulosic biomass into ethanol. Toward this task, the application of MS-based proteomics will not only provide key insights into which proteins regulate lignin-specific biological processes (for example, the monolignol biosynthesis pathway), but also insights that can be used to infer proteins that are implicated in the emergence of other bioenergy-relevant phenotypes (for example, enhanced plant growth).

1.4.2 Introduction to Populus

With the growing interest in the utilization of high biomass producing plant species, the *Populus* genus has become a research focal point and is considered to be among the major potential feedstocks for biofuel production[110-111]. Although other plants are being considered for use, only a few – including *Populus trichocarpa* – have a

sequenced genome and a range of genetic tools that enable genome-assisted crop improvement. In 2006, black cottonwood (*P. trichocarpa*)[112] became the first tree to have its genome fully sequenced and thus emerged as a model for tree genomics. *Populus* was chosen as the first tree to have its genome sequence because the genome is relatively compact, roughly 50 times smaller the genome of pine. Moreover, *P. trichocarpa*, and most *Populus* species in general, have rapid growth and reach reproductive maturity in as few as 4 years. The genome is of modest size, divided into 19 chromosomes, and is approximately four times larger than the first sequenced plant, *Arabidopsis thaliana*[113]. Analysis of the genetic organization of the genome has revealed two whole-genome duplication events, resulting in two-thirds of protein-coding genes sharing sequence similarity greater than 90%[112]. To date, the latest improved version of the poplar genome (v3 assembly) has 181 scaffolds that are greater than 50 kb in size, which represents 97.3% of the genome. In total, there are 41,335 loci that become transcribed into 73,013 protein-coding transcripts (http://www.phytozome.net/cgi-bin/gbrowse/poplar).

The *Populus* genus possesses tremendous genetic and thus phenotypic diversity - the genetic diversity across natural populations of *Populus* is extensive, in which a single nucleotide polymorphism occurs roughly every 200 base pairs[114]. In contrast to other plant models such as *Arabidopsis* and rice, which are predominately self-fertilizing and consequently maintain low levels of allelic polymorphism, the *Populus* genus is primarily composed of dioecious, self-incompatible woody plants. Obligate outcrossing combined with wind-pollination and prolonged reproductive life generates highly heterozygous populations with low levels of linkage disequilibrium[115]. This type of mating system results in high levels of gene flow and extensive nucleotide variability within and across *Populus*. With the advent of high-throughput sequencing, it has become feasible to sample the natural variation across a population, thereby enabling a more complete understanding of how genetic variations translate to phenotypic plasticity[116]. Therefore, widespread association efforts are being performed to reveal not only the genes, but also the natural variations underlying favorable traits[116]. By identifying the set of genes (i.e., proteins) and polymorphisms responsible for favorable traits, not only will we gain a better understanding of the biological pathways underlying plant growth and plant

adaptation, but also a necessary knowledge that can be leveraged towards improving *Populus* as a bioenergy feedstock.

1.4.3 Populus Research in the Systems Biology Era

With the completion of the *Populus* genome in 2006, numerous "omics"-based experimental strategies have offered scientists the ability to comprehensively explore and understand the complex network of genes, transcripts, proteins, and metabolites at a systems-level[117]. Since the release of the genome, research endeavors have propelled our understanding of not only biological processes related to growth and development, but also tree-specific traits such as long-term perennial growth, wood formation, and seasonality. The major set of functional genomics tools that were immediately developed and made available to study *Populus* were DNA and transcript microarrays. Due to their advanced technical maturity, microarrays provided the appropriate depth to expand our understanding of genes regulated during wood development[118], flower dormancy[119], and the genetic signatures that facilitate plant adaptation[120]. While genomic- and transcriptomic-based approaches certainly have provided an undeniable value towards answering tree-specific questions, critical informational gaps remain. Specifically, the ability to obtain deep and comprehensive protein measurements is currently lacking. In fact, the development and application of proteomics for plant-based research is still in its infancy.

So far, attempts to generate proteome data sets for *Populus* have failed to reach adequate proteome coverage. Unlike microbial proteomic measurements, which typically afford near complete proteome coverage, the application of MS-based proteomics to plant analysis has experienced slow progress. In fact, the most commonly used platform for plant proteomics has been two-dimensional gel electrophoresis (2-D PAGE) followed by protein sequencing via mass spectrometry. Although a number of plant proteomic studies published to date have used this platform to successfully map proteomes of various cells, tissues, and organs, these data sets do not constitute comprehensive assays[121-125]. More recently, on-line chromatographic mass spectrometry-based proteomics has dramatically extended the throughput and depth of protein identification in complex mixtures by interfacing multidimensional liquid chromatography with nano-electrospray tandem mass

spectrometry (2D-LC-MS/MS)[126-128]. Using this gel-free approach, bottom-up proteomics (analysis of proteolytic peptide mixtures) has started to provide detailed qualitative and quantitative observations of cellular metabolic activity for *Oryza sativa*, *Arabidopsis*, and *Populus*. Despite the progress being made, the performance features of contemporary methodology, technology, and bioinformatic tools used to generate the presence and amount of proteins have yet to provide reliable proteome coverage.

1.5 Overview of Dissertation

The research presented in this dissertation demonstrates how specific challenges posed by plant biology led to the advancement of mass spectrometry-based proteomics. The dissertation focuses on the development and application of a high-performance mass spectrometry-based technique for bioenergy research, in particular focusing on *Populus* proteomics, in an effort to understand how this information might be used to enhance the conversion of cellulose to biofuels. This dissertation outlines the pursuit of the most comprehensive, most accurate survey of the *Populus* proteome, with the following goals: (1) design a bioinformatic workflow that addresses a protein inference problem with respect to genetic duplications, (2) develop a method to maximize protein identifications, and (3) develop a mass spectrometry-based method for the identification of single amino acid polymorphisms. Chapter 2 of this dissertation provides experimental details related to each sample preparation method, experimental procedures, background related to the mass spectrometers used as well as their configuration parameters, and details related to the bioinformatic methods applied. Chapter 3 demonstrates the application of a pre-existing experimental strategy to generate a partial proteome map of *Populus* vascular tissue. In addition, Chapter 3 describes the level of intra-proteomic similarity in *Populus* and the development of a bioinformatic approach that resolves protein sequence redundancy. Chapter 4 describes the development of experimental procedures to overcome the challenges in plant cell lysis, protein extraction, and peptide sequencing. After addressing the bioinformatic and analytical challenges posed by *Populus*, Chapter 4 also highlights the precision and comprehensiveness of the optimized experimental strategy. Chapter 5 discusses the genetic diversity of *Populus* and how to identify

unexpected single amino acid polymorphisms in a proteome. Chapter 6 serves as a conclusion of the research presented in the dissertation as well as observations of the current state of mass-spectrometry-based proteomics and an outlook of the future. Herein are presented methods used and developed to tackle the notorious complications currently presented to the field of proteomics.

CHAPTER 2

METHODS, INSTRUMENTATION, AND BIOINFORMATICS

2.1 Methods

2.1.1 Populus Cell Lysis

In this chapter, the bottom-up proteomic experimental workflow will be outlined for all the research presented in this dissertation. For the research presented in Chapter 3, cells were obtained from clonally propagated stem cuttings of two *Populus* clones that were grown under standard cultural greenhouse conditions as previously outlined[128]. The two clones used were 'WV94', a *Populus deltoides* clone, and '717', a *P. tremula x alba* clone. Cuttings were allowed to grow under normal conditions for six months and then half of the trees in each clone were subjected to tension stress by bending the stem from the apical meristem to the mid stem. After two weeks, xylem and phloem tissue samples were collected from the upper (tension) and lower (opposite) sides of bent stems as well as erect control (normal) stems as previously described[128]. Next, six ramets per tissue type per genotype were pooled together for proteomic measurements. For cell lysis, the plant tissue was ground under liquid nitrogen using a mortar and pestle. For each growth condition, a 3 gram sample of ground tissue was suspended in 15 mL lysis buffer containing 125 mM Tris (pH 8.5), 10% glycerol, 50 mM DTT and 1 mM EDTA[129]. The suspension was vortexed twice for 30 seconds each time, then sonicated (Branson 185 sonifier, power setting of 40) on ice for three rounds of 30 seconds. Cellular debris was removed from the sample by centrifugation for 5 minutes at 1,200 x g, followed by centrifugation again for 10 minutes at 12,000 x g.

The architecture of plant cell walls provides resistance to chemical and biological degradation, thus requiring mechanical and detergent-based lysis for optimal proteome analysis. However, this criterion poses a major challenge for plant proteomic research using electrospray mass spectrometry, as detergent-containing solutions can impede enzymatic digestion and cause significant analyte suppression.

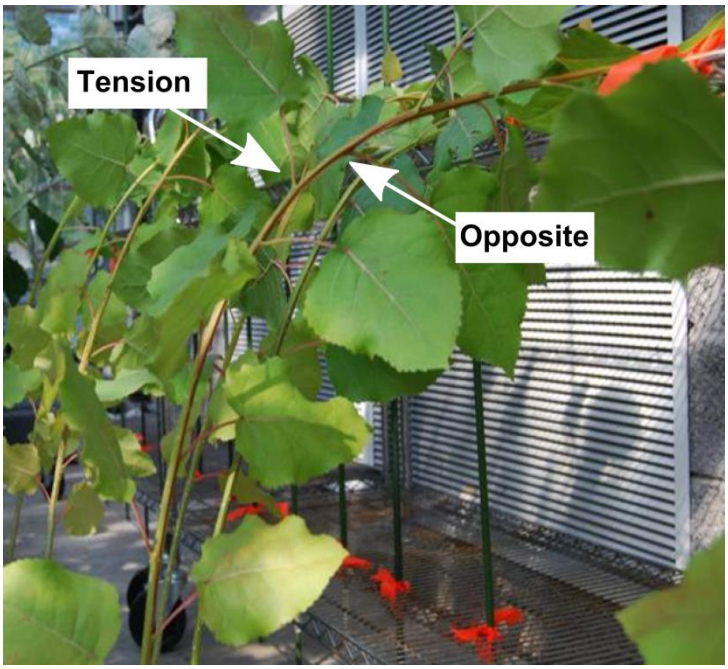


Figure 2.1. Illustration of *Populus* tree with tensional stress. Three types of stem tissue were harvested: 1) Tension, 2) opposite, and 3) Normal (i.e., no gravitational stress applied)

Therefore, most plant proteomic studies are forced to employ only mechanical disruption (*vide supra*). For the research presented in Chapters 4 and 5, we developed a detergent-based lysis approach for plant tissue in order to obtain a more comprehensive, a less biased proteome characterization well beyond that achievable with the pre-existing strategy employed in Chapter 3. For the research discussed in Chapters 4, 5, and 6, cells were obtained from the *P. tremula x alba* clone, '717', which was grown under standard greenhouse conditions. From these trees, mature fully expanded leaf including the petiole and midrib (leaf plastochronic index (LPI) 10-12) and young leaf including the petiole and midrib (LPI 4-6) samples, fine roots less than 2 mm in diameter and young photosynthetically active stem segments less than 5 mm in diameter were collected. Harvesting tissue samples from 6 month old trees afforded little biomass and thus confined the experimental design to only a single biological replicate per organ type. To reduce the effects of biological variation, tissues across 6 individual trees were pooled together for each organ type. For chapter 6, cells were obtained from two *P. trichocarpa* genotypes, 'DENA' and 'VNDL', which were grown under standard greenhouse conditions. From these trees, young leaf including the petiole and midrib (LPI 4-6) samples, fine roots less than 2 mm in diameter, and young photosynthetically active stem segments less than 5 mm in diameter were collected. Again, tissue was harvested from 6 individual ramets per genotype and pooled together for each sample tissue type to reduce the effects of biological variation. For cell lysis, leaf, root and stem tissues were each pulverized under liquid nitrogen using a mortar and pestle. For each biological sample, a 1.5 gram sample of ground tissue was suspended in sodium dodecyl sulfate (SDS) lysis buffer (4% SDS in 100 mM of Tris-HCl), boiled for 5 minutes, sonically disrupted (40% amplitude, 10 seconds pulse with 10 seconds rest, 2 minutes total pulse time), and boiled for an additional 5 minutes. Crude protein extract was pre-cleared via centrifugation at 4500 x g for 10 minutes.

Following each cell lysis procedure described in this dissertation, protein concentrations were determined using a bicinchoninic acid (BCA) protein assay (Pierce Biotechnology; Rockford, IL) – this is a critical step in proteomic studies because it

identifies the protein concentration obtained per lysis procedure, thereby allowing one to standardize the amount of protein being investigated[130-131].

2.1.2 Protein Extraction and Precipitation

For samples generated for Chapter 3, proteins were extracted through centrifugation at 100,000 x g for 1 hour, yielding a crude soluble protein fraction (cytosolic fraction) and a pellet (pellet fraction). Although this extraction strategy did provide sufficient protein recovery for downstream analysis, this approach does not take into consideration the molecular composition of plant tissue. In particular, plant material is more problematic because the cell wall contains complex storage polysaccharides, lipids, phenolic compounds, and a broad variety of secondary metabolites. As such, the prevalence of these compounds represents one of the most significant challenges associated with plant proteome analysis. Therefore, we implemented a protocol that improves protein recovery and removes contaminants via protein precipitation with trichloroacetic acid (TCA) and acetone. For samples generated for Chapters 4 through 6, lysed cell fractions were adjusted to 20% TCA, vortexed, and stored at -80° C freezer overnight. Precipitated proteins were pelleted by centrifugation at 21,000 g, 4° C for 15 minutes. Supernatants were discarded and 1 mL of cold acetone (-80°C) was added to each pellet to remove lipids and excess SDS. To expose the entire surface area of the protein pellet to acetone, each sample was vortexed. Next, pellets were centrifuged at 21,000 g, 4° C for 5 minutes. The supernatants were removed, and this process of washing precipitated proteins with cold acetone was repeated two more times. This washing process was followed by air-drying to remove the acetone and produce a dry pellet of proteins.

2.1.3 Protein Digestion

For the samples generated in Chapter 3, protein extracts were denatured and reduced with 6 M guanidine/10 mM dithiothreitol (DTT) for 1 hour at 60 °C. These denatured and reduced samples were diluted with 50 mM Tris-HCL/10 mM CaCl2 (pH 7.6) to reduce the guanidine concentration to 1 M. Proteins were enzymatically digested into peptides with 1:100 (w/w) sequencing-grade trypsin (Promega; Madison, WI) at

37°C overnight, followed by a second addition of the same amount of trypsin and incubation for an additional 4 hour at 37°C. Centrifugation (3000 g for 10 minutes) was performed to remove any remaining cellular debris from solution. Digested peptides were desalted off-line using C_{18} solid phase extraction via SepPak Plus C18 cartridges (Waters, Milford, MA), eluting peptides using 100% acetonitrile (ACN). Peptide mixtures were concentrated using vacuum centrifugation (SpeedVac, Savant Instruments, Holbrook NY), bringing the final volume to ~500 µL. Peptide concentrations for each sample were determined using a bicinchoninic acid (BCA) protein assay.

For samples generated for Chapters 4 through 6, each TCA/acetone precipitated pellet was brought up in 250 µL of 8 M urea in 100 mM Tris buffer and sonicated for 2 minutes at 20% amplitude (5 seconds on and 10 seconds off) to fully solubilize the protein pellet. Next, proteins were allowed to denature at room temperature for 30 minutes. Denatured proteins were reduced with DTT (5 mM) and cysteine residues were blocked with iodoacetamide (20 mM) to prevent reformation of disulfide linkages. Samples were enzymatically digested via two aliquots of sequencing-grade trypsin (Promega, 1:75 [w/w]) at two different sample dilutions, 4 M urea (overnight) and 2 M urea (4 hours). Following digestion, peptide mixtures were adjusted to 200 mM NaCl, 0.1% formic acid and filtered through a 10 kDa cutoff spin column filter (Vivaspin 2, GE Health; Littleton, MA) to remove any remaining cellular debris as well as under-digested proteins. The peptide-enriched flow through was then quantified by BCA assay and stored at −80°C until MS analysis.

2.1.4 Liquid Chromatography

For all bottom-up proteomic experiments discussed in this dissertation, high-performance liquid chromatography (HPLC) was employed to separate peptide mixtures prior to MS analysis. Specifically, all proteome analyses were performed using multidimensional protein identification technology (MudPIT). First described by Yates and colleagues[132], the method couples two-dimensional - strong cation-exchange (SCX) resin and reversed-phase resin – liquid chromatography (LC) in a microcapillary chromatographic column. Using this approach, a fused silica microcapillary column (150 µm inner-diameter and 360 µm outer-diameter; Polymicro Technologies; Phoenix, AZ)

was prepared by pressure-loading ~3 cm of SCX resin (Luna 5 µm particle size; 100 Å pore size; Phenomenex; Torrance, CA) followed by ~3 cm of C_{18} reversed-phase resin (Aqua 5 µm particle size; 125 Å pore size; Phenomenex, Torrance, CA) - for clarity, these columns will be referred to as *back-columns*. Each prepared back-column was washed for ~5 minutes with Solvent B [30% HPLC grade water, 70% acetonitrile (ACN), 0.1% formic acid (FA)] and equilibrated for ~5 minutes with Solvent A [95% HPLC grade water, 5% ACN, 0.1% FA]. For each biological sample, volumes corresponding to a specified peptide concentration (25 to 100 µg) were pressure-loaded directly onto a back-column. After pressure-loading a sample, each back-column was washed offline for ~45 minutes across a linear gradient from 100% Solvent A to 50% Solvent B to remove any salt or excess SDS.

Next, back-columns were interfaced with a quaternary HPLC pump (Ultimate 3000 HPLC; Dionex; Sunnyvale, CA) that supplied three buffers through a three-way connecting PEEK MicoTee (Upchurch Scientific; Oak Harbor, WA). The buffers used were: (i) Solvent A (95% HPLC grade water, 5% ACN, and 0.1% FA), Solvent B (30% HPLC grade water, 70% ACN, and 0.1% FA), and (iii) Solvent C (500 mM ammonium acetate in Solvent A). The MudPIT microcapillary plumbing system (Figure 2.2A) splits the flow from the HPLC to either the mass spectrometer or to waste. This design served two purposes: the majority of salt was directed to waste and a 300-400 nL/min flow directly to the nano-ESI source of the MS maintained. With this setup, back-columns were connected to the branching point of the MicroTee directing flow to the mass spectrometer. The back-column was then placed in-line with an in-house pulled nanospray emitter (100 µm inner-diameter and 360 µm outer-diameter) packed with 15 cm of reversed-phase resin material – for clarity, these columns will be referred to as *front-columns*. The back-and front-columns were connected via a PEEK union and 0.5 µm inline filter. To direct flow to waste, a small piece (~6 inches) of fused silica bridged the first MicroTee to a second MicroTee. On the second MicroTee, one branching point was attached to a gold electrode – this provided the voltage supply (1-6 kV) necessary for electrospray - and a second branching point was connected to fused silica that served as a waste-line. In addition, the length of the waste-line could be adjusted to regulate the flow-

rate at the front-column emitter – typically, a length of ~2.5 feet would create a back pressure of 70-80 bar, resulting in a flow rate of approximately 400 nL/min.

The primary strength of the MudPIT approach is the orthogonality of the two chromatographic phases: peptides are selectively displaced, by their isoelectric point (i.e., charge), from the SCX resin by controlling the salt concentration (i.e., Solvent C), and then peptides are separated across the reversed-phased resin based on their "stickiness" (i.e., determined by a peptide's hydrophobicity) via a linear organic gradient (i.e., Solvent B)[133-134]. These LC separations were performed on-line by attaching the MudPIT system to a nanospray source (Proxeon, Denmark), which is mounted in front of the mass spectrometer. Due to the size and complexity of the peptide mixtures being analyzed, we employed a stepwise separation scheme that consisted of 11 fractionation steps, each lasting ~two hours (Figure 2.2A). In a single step, the conjoined back- and front- column were first washed with Solvent A (100%), and then a short segment of Solvent C was applied, followed by a long gradient of increasing Solvent B (0% to 50%). From steps 1 to 10, the concentration of Solvent C was increased in small increments from 5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20%, 25%, 35%, to 50%. In the last step, the separation scheme in step 10 is repeated, with the exception that Solvent B gradient reaches 100%. While peptides were being separated and eluted from the front-column, they were subsequently ionized via electrospray and introduced into the mass spectrometer (*vide infra*).

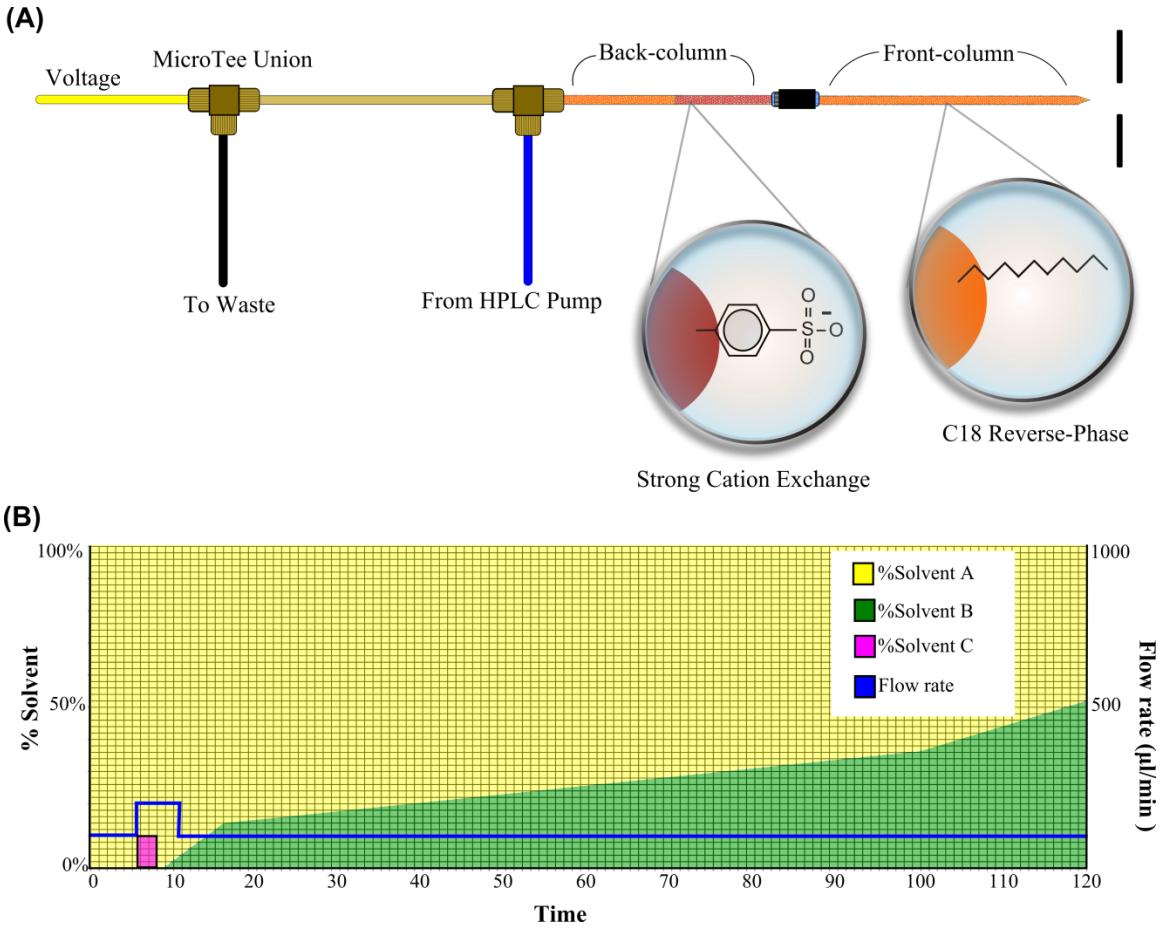


Figure 2.2. Illustration of a (A) MudPIT experimental setup and a (B) multi-step HPLC gradient.

2.2 Instrumentation

2.2.1 Analytical Figures of Merit

Mass spectrometer instruments typically require the following three parts: the ion source, the mass analyzer, and the detector. Over the past two decades, various technological developments and adaptations of each component have made MS one of the most versatile tools to characterize proteins. As such, there are many possible ways to ionize, analyze, and detect peptides. Although there is no single MS instrument configuration that is superior to all others, only a relative few types of MS instrumentation dominate the field of proteomics. When choosing a mass spectrometer instrument for a specific application, proper selection becomes a balancing act that hinges on the following analytical figures of merit[135]: mass resolving power (i.e., the ability to distinguish between ions of different *m/z* ratios and is obtained by calculating the full width at half-height of a single well-resolved peak), mass accuracy (i.e., the ratio of the *m/z* measurement error divided by the true *m/z* and is usually stated in terms of parts per million), mass range (i.e., the range of *m/z* ratios amenable to analyze by a given mass analyzer), dynamic range (i.e., a measure of the detection range of a detector and is calculated as the ratio of the largest to smallest detectable signal), precision (i.e., the reproducibility with which ion abundances can be determined), duty cycle (i.e., the fraction of time that the instrument is collecting data), sensitivity (i.e., the inverse of the ratio obtained by dividing the signal level of the largest peak in a spectrum by the signal level of the background at one *m/z* value higher or lower) , and speed (i.e., the number of spectra per unit time that can be collected). For example, routine peptide identifications are mostly performed with ion traps instruments that are very sensitive and fast, but they have lower resolving power. In this scenario, the relative importance of resolving power is not as significant as it might be for another application, such as large intact protein measurements. In top-down proteomics, not only is high mass resolution (>100,000) an absolute requirement for isotope resolution of all protein charge states, but high mass accuracy is also a necessity for unambiguous identifications of unmodified and modified proteins by database searching. As a key a focal point of experimental design, the performance characteristics of each instrument will be highlighted.

2.2.1 Ionization Sources

At present, the most common method to generate peptide ions is via protonation (or deprotonation). Since peptides contain basic residues that readily accept protons, they are capable of forming stable cations (i.e., positively charged ions) – a single proton added to a peptide produces a net charge of +1. Currently, protonation of peptides is most commonly achieved using the electrospray ionization (ESI) process, which is considered well understood (Figure 2.3). The first suitable ESI-MS source was designed by Fenn and colleagues in the 1980s[19]. Without any voltages applied to an ion source, a droplet forms at the end of the front-column emitter – a small (micron-sized) droplet can contain thousands of peptide analytes. When a very high voltage (1-6kV) is applied to the emitter, the charged analytes (i.e., protonated peptides) are repelled by the high-voltage (of the same polarity), forcing the droplet to deform into a cone-shaped spray. First described by Sir Geoffrey Taylor[136], two combating physical effects (i.e., surface tension and Coulomb repulsion) cause the liquid solution to form the cone-shaped (i.e., Taylor cone) spray at the end of the emitter. When droplets form, the surface tension force tries to retain a spherical shape and distribute charges across the surface to minimize the potential energy[137]. While the surface tension force tries to retain the spherical shape of the droplet, the Coulomb force of repulsion between like charges tries to break down the spherical shape of the droplet[138]. As droplets traverse the space between the emitter and the heated capillary (for orientation purposes, the heated capillary is often 1-3 cm from the emitter), their solvent begins to evaporate. When solvent molecules leave, the charge density at the surface of the droplets begins to increase. This process occurs repeatedly to generate smaller and smaller droplets until a droplet reaches the point (Rayleigh limit) where the surface tension on the charged droplet can no longer sustain the Coulomb force of repulsion[139]. At this point, "Coulomb explosion" occurs and forms even smaller droplets[140]. After the process of solvent evaporation and Coulomb explosion occur several times, the gas-phase charged peptide analyte is formed[141]. Once the gas-phase ions have been produced, the ions become attracted to the entrance of the mass spectrometer (i.e., the heated capillary) due to the high opposite voltage of the entrance.

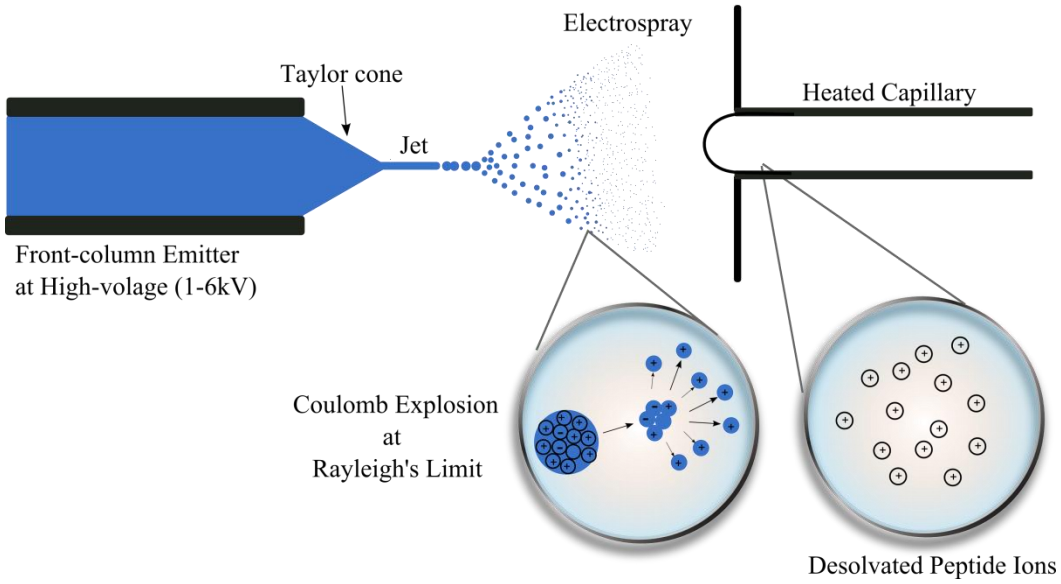


Figure 2.3. Illustration of the electrospray ionization process.

As the interface between atmospheric pressure and high vacuum, the heated capillary also helps remove remaining traces of solvent – the heated capillary is maintained at temperatures between 200-275 °C.

Overall, the entire electrospray process takes around a few micro-seconds and is conducive to the formation of multiply charged peptides. This is an important analytical advantage for MS instrumentation. Since MS measures the *m/z* value, this phenomenon makes it possible to observe very large molecules with an instrument that has a limited mass range – for an ion trapping instrument the mass range is typically from 150 to 2000 *m/z*. Thus, a *m/z* value of a singly charged protein that has a molecular weight of 10 kDa may exceed the mass range of most instruments; however, by increasing the number of charges on the protein, for example to +10, the *m/z* ratio of the protein is decreased to a level that is measurable. Moreover, the presence of multiple charge states of the same molecule supplies multiple measurements of the same molecular species. Despite these advantages, ESI is not without challenges. Specifically, some analytes in a solvent droplet can change the efficiency of other analytes to become ionized. For example, it has been shown that for larger molecular ions, hydrophobic molecules have higher ionization efficiency than hydrophilic ones[142]. The presence of ion suppression makes it difficult to compare signals between different peptides - especially for lower abundant peptides species, where the signal-to-noise level compromises the precision and accuracy of the measurement. Although ion suppression cannot be completely avoided, some adjustments can be made to counter ion suppression, for example modifications to chromatographic conditions can minimize peptide co-elution.

Currently, peptides are most commonly ionized using nano-ESI (operating at a flow rate of nL/minute)[143]. While conventional ESI sources produce primary charged droplets of 1-2 µm in diameter, the nanospray version of ionization produces smaller droplets that are 100-200 nm in diameter, which have several desirable analytical properties: nano-ESI enables longer analyte signals for analysis, better desolvation of ions, and less consumption of the sample mixture[144]. In addition, nano-ESI sources are positioned closer to the entrance of the mass analyzer, therefore ion transmission is much more efficient, with overall ionization efficiencies ~500 times higher than an electrospray

source[144]. Collectively, nano-ESI achieves higher sensitivity than the conventional macro-ESI sources – low-attomolar range versus femtomolar range, respectively[145].

Since ESI involves the continuous introduction of a liquid, it is easily coupled with HPLC. Hence, peptide ionization was achieved by interfacing the MudPIT platform with a nano-ESI source (Proxeon; Odense, Denmark). Although many solvents can be used in ESI, it is important to stress that ESI solvents are most effective when a volatile organic solvent is present and the solvent is capable of donating a proton - for this reason, each solvent used for HPLC consisted of 0.1% formic acid and some percentage of acetonitrile. Once peptides were subjected to the gas phase, they were electrostatically transmitted to the mass analyzer region of the mass spectrometer, where peptide ions were sorted and separated according to their *m/z* values.

2.2.2 Analyzers and Detectors

The second major component of the mass spectrometer is the mass analyzer, which is responsible for separating the different types of ions (*m/z* values) contained in an ion beam. The heart of every mass spectrometer, the mass analyzer, is a region that has numerous configurations. Despite the remarkably diverse types and arrangements available, the most common type is the ion trap mass analyzer. As discussed in Chapter 1, the ion trap is a tandem-in-time mass analyzer, which means that ion accumulation, selection, and dissociation all occurs within the same analyzer but at different times. For Chapter 3, the linear trapping quadrupole (LTQ) XL mass spectrometer (Thermo Scientific, Waltham, MA) was used for peptide sequencing. For Chapter 4 and 5, the next generation LTQ Velos mass spectrometer (Thermo Scientific, Waltham, MA) was used for peptide sequencing. For Chapter 6, the hybrid LTQ-Orbitrap Pro (Thermo Scientific, Waltham, MA) was employed for the sequencing of modified peptides. The schematics of each mass analyzer and their performance metrics are depicted in Figure 2.4. The performance comparisons of the mass spectrometer instruments can be seen in Table 2.1.

In general, the single linear ion trap (LIT) mass analyzers are two-dimensional rectangular ion storage devices, composed of an array of four rods with a space down the central axis[146-147]. Ion motion through the central axis in the quadrupolar arrangement is established by manipulating the electrical surfaces of the rods.

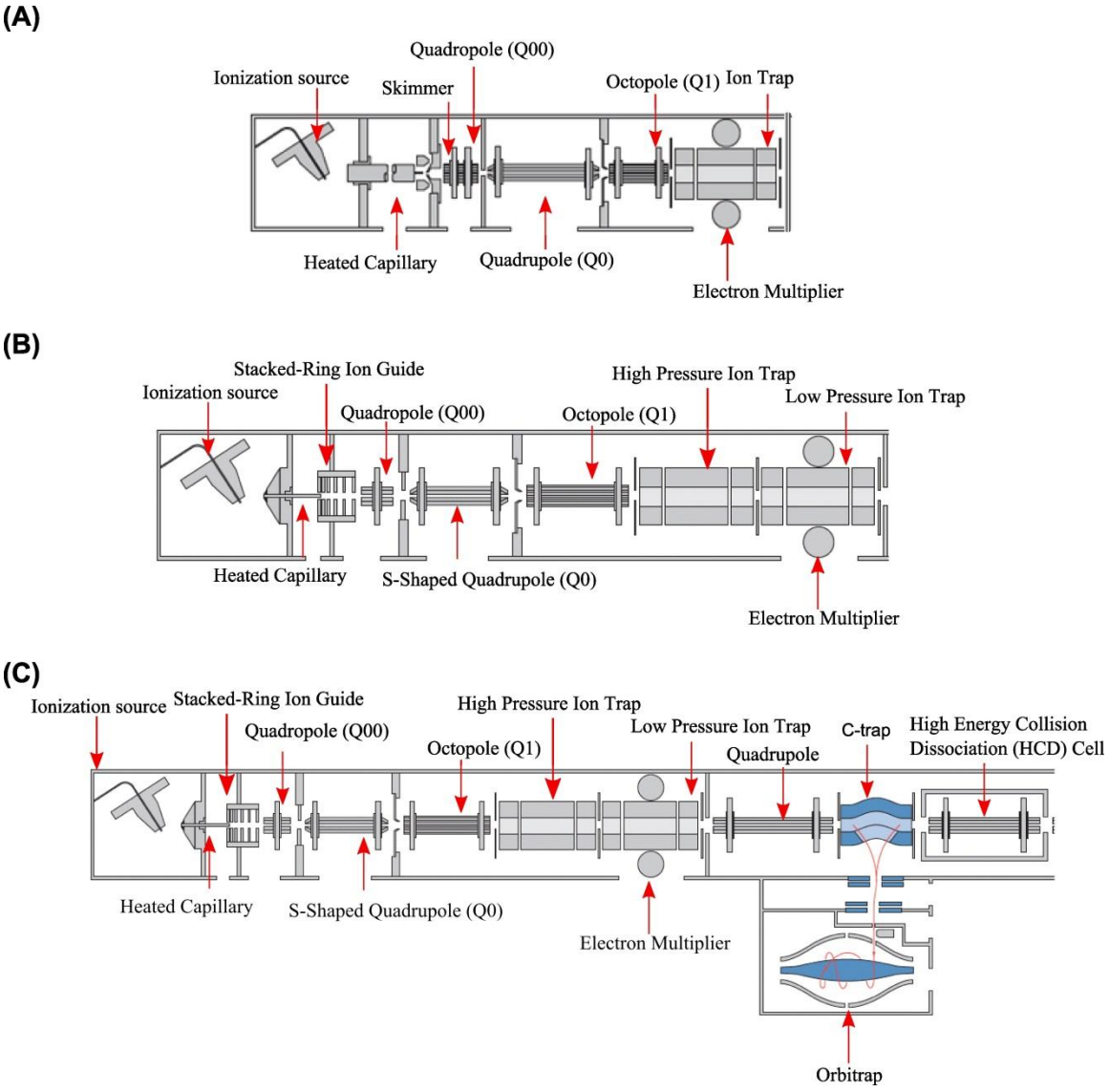


Figure 2.4. Schematics of the (A) LTQ XL mass spectrometer, (B) LTQ Velos mass spectrometer, and (C) the LTQ-Orbitrap Pro mass spectrometer. These illustrations were from the previous schematics[148-150].

Table 2.1. Performance metrics of instrumentation.

Instrument	Mass Resolution	Mass Accuracy	Sensitivity	Dynamic Range	Data Acquisition Speed	Duty-Cycle
LTQ XL	2000-10,000	100ppm	Femtomole	3-4	Moderate	Moderate
LTQ Velos	2000-25,000	1ppm	Attomole-Femtomole	5-6	Fast	Very High
LTQ-Oribtrap Pro	15,000-400,000	<1ppm	Attomole-Femtomole	5-6	Moderate	Moderate

* The dynamic range reported is the order of magnitude of the ratio between the highest and least abundant ion signal. The relative data acquisition speed represents the number of scans collected per second at a specific mass resolution. The speeds are the following: ~2 scans/second (LTQ XL; resolution = 2000), ~4 scans/second (LTQ Velos; resolution = 4000), and ~3.5 scans/second (LTQ-Orbitrap Pro; resolution = 30,000).

Within an ion trap, ions are confined radially (x- and y-direction) by a two-dimensional radio frequency (RF) potential and axially (z-direction) by a direct current (DC) potential applied to the front and back end caps (electrodes), controlling the longitudinal trajectory of ions. In addition, a constant DC voltage and a supplemental alternating current (AC) voltage are applied to the rods to regulate the axial trapping and radial excitation, respectively. Together, this creates a harmonic electrostatic field that can be altered for ion trapping, manipulation of ion trajectories, and *m/z*-selective ion ejection. The stability of an ion inside an ion trap can be defined by the Mathieu equations:

$$a_x = -a_y = (4eU) / (m^2 r_0^2) \qquad \text{Equation 2.2.2.1}$$

$$q_x = -q_y = (4zeV) / (m^2 r_0^2) \qquad \text{Equation 2.2.2.2}$$

where the *a* variable is related to the DC voltage, *q* is related to RF voltage, *e* is charge state of the ion, *U* is the amplitude of the DC voltage, *V* is the amplitude of the RF voltage, *m* is the mass of the ion, and r_0 is the distance from the z-axis. Examination of these equations shows that stable ion trajectory (i.e., the ion follows a trajectory that avoids collisions with the surfaces of the trap) is a function of both mass and charge[151]. By solving the *a* and *q* parameters, the stable x-z and y-z ion trajectories can be computed. These mathematical transformations can be summarized by the stability diagram illustrated in Figure2.5. Here, the six-dimensional problem is reduced to a two-dimensional plot. In order to trap ions, there must be quadrupolar stability in both dimensions. This is achieved by applying a constant DC voltage (axial trapping field), where $a = 0$, and applying a low auxiliary AC and RF voltages (radial trapping field). Under this scenario, ions map to the x-axis of the stability diagram – the x-coordinate of the ion is dictated by the *m/z* of the ion, where the ions are positioned left to right from the heaviest to the lightest, respectively. As explained in Equation 2.2.2.2, the mass of an ion is inversely proportional to *q*; therefore, by increasing the RF magnitude from low to high, ions of increasing mass will sequentially take on unstable trajectories in the radial dimension. The ability to "scan out" ions by mass provides a way for linear ion traps to selectively isolate, activate, and eject ions to the detector – where the ejected ions are converted to an electronic signal (i.e., a mass spectrum).

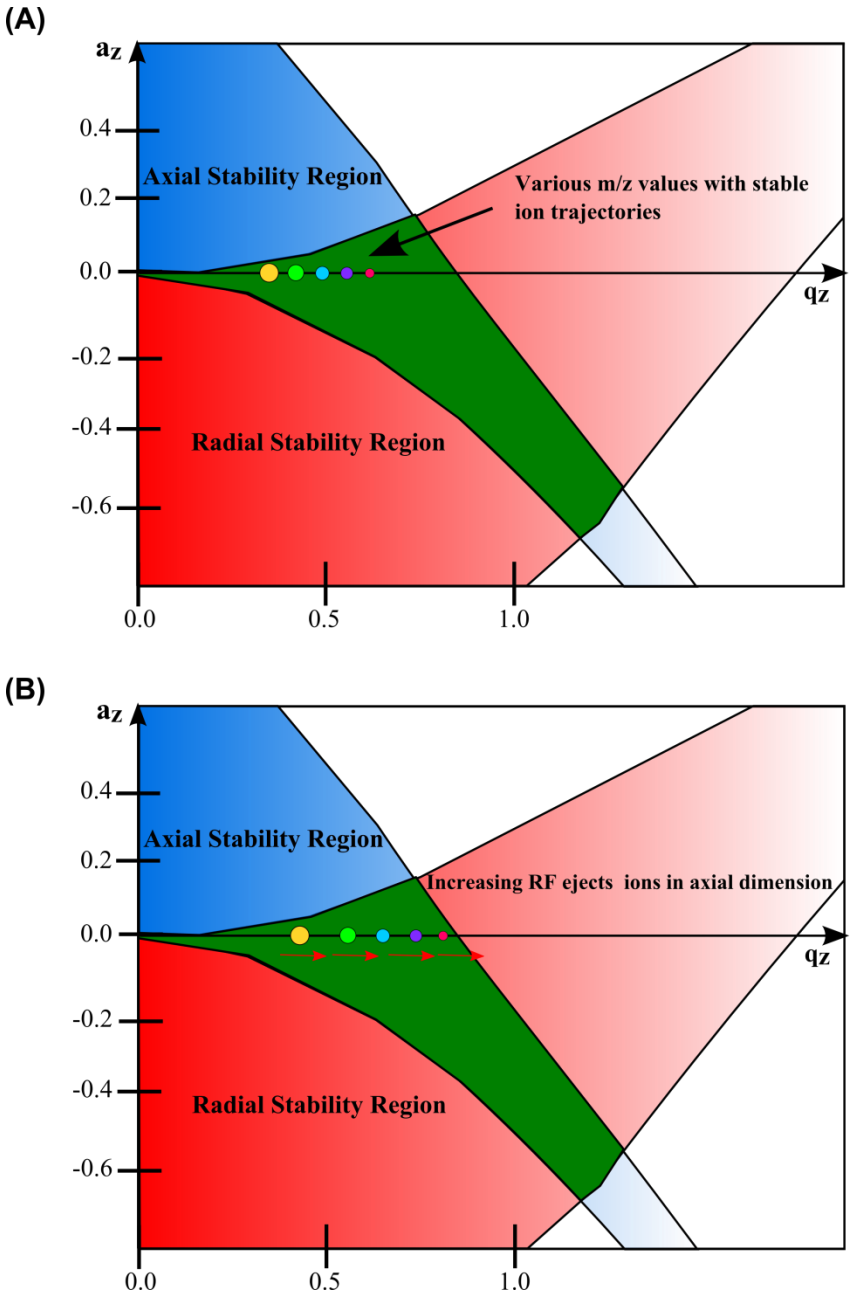


Figure 2.5. Illustration of the quadrupolar stability diagram.. Depending on the various *a* (y-coordinate) and *q* (x-coordinate) values, the ions can be (A) trapped or (B) ejected from the analyzer. The blue region represents an ion *m/z* value that has stability in the axial dimension of the trap and the red region represents stability in the radial dimension of the trap. The green region represents a region where ion *m/z* values have stability in the axial and radial dimension. These illustrations were adapted from a previous figure[152].

The most widely used LIT detection system consists of a continuous-dynode electron multiplier (i.e., channeltron), and an electrometer[152]. After positively charged ions are ejected from the sides of the LIT analyzers, they are accelerated to a high velocity by holding the conversion dynode at a relatively high negative potential – this is done to improve detection efficiency. As these ions strike the curved surface of the dynode cup, the impact yields the emission of several electrons. These electrons pass further into the electron multiplier, again striking the wall. This snowball effect is repeated multiple times, amplifying the initial signal by producing more and more electrons[153]. At the exit of the detector, the ion current is measured and converted into a voltage via the electrometer and converted to an intensity value (the y-axis on a mass spectrum).

Recently, the innovative dual-pressure linear ion trap mass spectrometer (LTQ Velos)[150] was developed to provide improvements in acquisition speed, sensitivity, and ion isolation and fragmentation. Rather than using only a single ion trap, the dual-pressure LIT system consists of two trapping devices separated by a single aperture lens to allow differential pumping between the traps. Relative to the LTQ XL, the first trap is held at a higher pressure (~5 10^{-3} Torr), providing improvements to ion trapping, ion isolation, and ion fragmentation efficiencies. The second trap, which is used to scan out ions for detection, is held at a lower pressure (~5 10^{-4} Torr), which increases the acquisition speed and reduces fragmentation while ions are being ejected from the trap. This type of sensitivity, dynamic range, and data acquisition speed becomes critical for the analyses of complex mixtures, such as *Populus*.

In contrast to the linear ion trap mass analyzers, the Orbitrap mass analyzer[154] is used as part of a tandem-in-space mass spectrometer that has a LIT upstream. Unlike the LIT analyzers, there is no RF to hold ions. Instead, moving ions are trapped in a constant electric field that is established by two opposing electrodes - the analyzer consists of an inner axial (spindle-shaped) electrode and an outer coaxial (barrel-shaped) electrode[155]. When ions are injected into the Orbitrap, the ions become attracted to the inner electrode through electrostatic attractions. The ions begin follow a circular orbit (hence, the name) because the electrostatic attraction towards the inner electrode is compensated by a

centrifugal force that arises from the initial velocity of ions. While the ions oscillate around the spindle, they also traverse the z-axis (i.e. longitudinal direction). The ion motion in the z-direction can be described as a harmonic oscillator and the *m/z* ratio of the ions is simply related to the frequency of the oscillation along the z-axis:

$$\omega = [(z/m)k]1/2 \qquad \text{Equation 2.2.2.3}$$

where ω is the frequency of axial oscillation, z is the ion's charge, m is the ion's mass, and k is field curvature. For the Orbitrap analyzer, ion detection is performed by broadband image current detection, followed by a fast Fourier transform algorithm to convert each time-domain signal into their respective *m/z* signals. Such measurements achieve very high resolution (~400k) - similar to those achievable by FT-ICR instruments - and surpass the resolution obtainable by the ion trap instruments[156]. Due to the high resolving power, the Orbitrap can achieve highly accurate mass measurements (<1 part per million). The level of resolving power and mass accuracy becomes critical for the analyses of protein or peptide modifications, where a high level of discrimination power is required.

2.2.4 Data-dependent Acquisition

In a conventional LC-MS/MS experiment, thousands of peptide species are co-eluting at any given time, especially for complex proteomes such as plants. Therefore, in any given MS^1 (full scan) there will be thousands of *m/z* ratios at varying signal intensities. Despite advancements in mass spectrometer instrumentation, the number of ions surveyed in a single full scan significantly exceeds current duty cycle (i.e., the fraction of time that the instrument is collecting data) capabilities[157]. Since the primary goal of bottom-up proteomics is to identify as many peptides as possible in LC-MS/MS runs, a sophisticated sampling strategy is required to maximize the sequencing efficiency of the mass analyzer. Today, the most commonly used strategy applies a data-dependent acquisition (DDA) method, which uses information collected in each full scan (MS^1) to selectively trigger the subsequent MS/MS experiment[158]. In particular, the mass spectrometer schedules peptide fragmentation events based on the peptide mass and intensity information from each full scan (MS1). In addition, the mass spectrometer

measurements vary from scan to scan based on the information acquired in previous scans.

The most widely used DDA scheme is a "topN" method in which each full MS scan is followed by up to N precursor isolation and fragmentation events[159]. To select a precursor for fragmentation, the software controlling the instrument sorts the precursor ions detected in each MS scan by intensity and then applies user-specified criteria such as minimum signal intensity, charge state, and dynamic exclusion (i.e., the avoidance of already fragmented precursors). Although all these parameters can affect data quality, dynamic exclusion is the most important for complex mixtures[160-161]. When dynamic exclusion is not enabled, the mass spectrometer continuously samples the top N ions. As a result, the mass spectrometer only samples the high abundance ions that dominate a full scan. Obviously, this would severely limit the depth of the proteome measurement. On the other hand, when dynamic exclusion is applied, the mass spectrometer will not repeat fragmentation on the same precursor ion. Once a precursor has been fragmented, it is put onto a dynamic exclusion list so that the mass spectrometer can trigger MS/MS events on other, usually less abundant, ions in each full scan. Therefore, both high- and low-abundance co-eluting peptides have a chance to undergo fragmentation.

For the research presented in Chapter 3, mass spectra were collected in a data-dependent "top5" mode: minimum precursor ions signal intensity of 1000, CID (35% energy) was used for fragmentation, and dynamic exclusion was enabled (max exclusion list size of 100 *m/z* values, an exclusion mass width of 1.5 Da, a repeat count of 1, and exclusion duration of 3 minutes). For the research presented in Chapter 4 and 5, mass spectra were collected in a data-dependent "top10" mode: a minimum precursor ion signal intensity of 1000, CID (35% energy) was used for fragmentation, and dynamic exclusion was enabled (max exclusion list size of 500 *m/z* values, an exclusion mass width of 1.5 Da, a repeat count of 1, and exclusion duration of 60 seconds). For Chapter 6, mass spectra were acquired in a data-dependent "top20" mode: a minimum precursor ion signal intensity of 1000, CID (35% energy) or HCD (40%) was used for fragmentation, and dynamic exclusion was enabled (max exclusion list size of 500 *m/z* values, an exclusion mass width of 0.2 Da, a repeat count of 1, and exclusion duration of

60 seconds). It is important to highlight that each DDA method was adapted to account for increases in instrument performance metrics.

2.3 Bioinformatics

2.3.1 Peptide Sequencing

Once MS data was collected, the MS/MS spectra were analyzed by database searching algorithms to automatically match experimental spectra to peptide sequences in a protein sequence database. For Chapter 3 and 4 measurements, the output files (Thermo .RAW) were searched with the SEQUEST database algorithm[42]. In brief, the algorithm performs four major steps: spectrum preprocessing, searching, preliminary scoring, and cross-correlation analysis. In the first step, each MS/MS spectrum is preprocessed by removing all but the top two hundred most abundant *m/z* values, and thereby removes noise peaks to improve the overall search performance and accuracy. Next, the algorithm searches each spectrum against peptide sequences (protease specific) from the user-supplied database of protein sequences. During the searching process, candidate peptide sequences are culled from the database by applying a few simple filters. The first filter used in identifying plausible sequence matches is the precursor ion mass (within a user-specified mass tolerance). The choice of precursor mass tolerance is dependent on the mass accuracy of the instrument used to acquire the data. Therefore, high mass accuracy instruments can enforce a more stringent mass tolerance to restrict the number of candidate peptide sequences being compared to each spectrum, and thus improving the overall quality of the data set. On the other hand, if the mass tolerance filter was too widely defined, the predicted peaks have a higher chance of matching to random peaks in the spectrum. In this scenario, the peak matching process is more likely going to generate false positive matches. Once a set of candidate sequences have been defined, each peptide sequence is then converted into a virtual spectrum (i.e., a list of predicted *m/z* values for fragment ions) and scored. During the scoring step, three factors are combined to produce a preliminary score: i) the summed intensity of matched ions is determined, ii) the continuity of each sequence is evaluated (additional scoring weight is given for successive fragment ions), and iii) the percentage of ions found versus those expected is

calculated. Using the preliminary score as a filter, a cross-correlation score is computed for the top five hundred candidate peptide sequences. The cross-correlation (XCorr) score represents an average of differences between the *m/z* values in the observed and virtual spectrum. As part of the cross correlation analysis, an additional score (DeltCn), is calculated to measure the scoring difference between the lowest ranked peptide scores and the XCorr value of the best match. That is, this score provides an indication of how well SEQUEST could distinguish the top peptide-spectrum match (PSM) compared to the second-best PSM. For example, a high DeltCn of the second ranked PSM means that there is a high likelihood that the top ranked match is correct.

For Chapter 3 database searching, experimental MS/MS spectra were compared to theoretical tryptic peptide sequences generated from a database containing (1) the protein database of *P. trichocarpa* (v2.0; 45,778 proteins), (2) predicted small proteins (20,565; 10–200 amino acids in length), and (3) common contaminant proteins (i.e., bovine trypsin and human keratin). A decoy database, consisting of the reversed sequences of the target database, was appended in order to determine the false-discovery rate (FDR) for protein identifications. Using this protein database, peptide fragmentation spectra (MS/MS) were assigned peptide sequences with the SEQUEST algorithm v.27, employing the following parameters: ≤ 4 missed tryptic cleavages allowed, precursor ion mass tolerance of 3.0 *m/z* units, fragment mass tolerance of 0.5 *m/z* units. For Chapters 4 and 5 database searching, experimental MS/MS spectra were compared to theoretical tryptic peptide sequences generated from a FASTA database containing (1) the full protein complement of *P. trichocarpa* (v2.2, released in 2011; containing 45,778 proteins), (2) mitochondria and chloroplast proteins, and (3) common contaminant proteins (i.e., bovine trypsin and human keratin). A decoy database, consisting of the reversed sequences of the target database, was appended in order to discern the false-discovery rate (FDR) at the peptide level. Using this protein database, peptide fragmentation spectra (MS/MS) were assigned peptide sequences with the SEQUEST algorithm v.27, employing the following parameters: ≤ 4 missed tryptic cleavages allowed, a parent ion mass tolerance of 3.0 *m/z* units, a fragment mass tolerance of 0.5 *m/z* units, and a static modification on cysteine (iodoacetamide; +57 Da).

For Chapter 6, the output files (Thermo .RAW) were searched with two different algorithms: MyriMatch[63] and TagRecon[162]. For traditional database searching, the MyriMatch algorithm was employed. The entire process has three major steps: spectrum preprocessing, searching, and scoring. Similar to SEQUEST, the MyriMatch algorithm attempts to simplify each MS/MS spectrum by removing noise peaks; however, whereas SEQUEST arbitrarily chooses a number of ions to retain, MyriMatch has a tunable preprocessing step that ranks the ions by their intensity and retains only the top N% of ions in each scan, where N is by default 98. Similar to SEQUEST, the algorithm searches each spectrum against peptide sequences (protease specific) from the user-supplied database of protein sequences, using the *m/z* tolerance to generate candidate peptides. Using basic fragmentation rules, a virtual spectrum is created for each candidate peptide sequence, listing *m/z* positions at which it expects to observe fragment ions. In contrast to SEQUEST, MyriMatch uses a sophisticated scoring system that is statistically derived, making it more interpretable. For each experimental spectrum, the software examines each *m/z* location and computes two probabilistic scores: an intensity-based MVH score and a mass error-based mzFidelity score. To compute the MVH score, peaks in the experimental spectra are first separated into three intensity classes (high, medium, and low). Importantly, these classes differ in the number of peaks they hold: the highest-intensity class will be sparsely populated and the low-intensity class will be more populous. For each predicted *m/z* value, the corresponding location in the experimental spectrum is tested to determine whether or not a peak match occurs, and if so, what intensity class it falls in. To compute the probability of whether or not this match occurs by random chance, MyriMatch employs a multivariate hypergeometric (MVH) distribution. Since the high-intensity class contains very few peaks, matching a peak to this lightly populated class will contribute more to the peptide score because it is highly unlikely to occur solely by chance. To compute the mzFidelity score, a similar distribution is created. Rather than using intensity, the fragment mass error distribution is determined to compute the probability of a fragment peak matching by random chance within three difference classes. In general, fragment ion peaks are rewarded for their proximity to their predicted *m/z* value. After both scores are calculated for each candidate

sequence, the algorithm uses the MVH scores to rank the candidate sequences and, when needed, the mzFidelity score serves as a tie-breaker. Finally, an Xcorr value is computed to independently validate the best PSM ranked by the MVH score.

For Chapter 4 and 5 database searching, experimental MS/MS spectra were compared to theoretical tryptic peptide sequences generated from a FASTA database containing (1) the full protein complement of *P. trichocarpa* (v3, released in 2012;73,013), (2) mitochondria and chloroplast proteins, and (3) common contaminant proteins (i.e., bovine trypsin, human keratin, etc.). A decoy database, consisting of the reversed sequences of the target database, was appended in order to discern the false-discovery rate (FDR) at the peptide level. For standard database searching, the peptide fragmentation spectra (MS/MS) were searched with MyriMatch algorithm v2.1, employing the following parameters: infinity tryptic cleavages allowed, an average parent ion mass tolerance of 1.5 *m/z* units or a monoisotopic precursor mass tolerance of 10 ppm (only when isotopic resolution was obtained), an average fragment mass tolerance of 0.5 *m/z* units, a total ion current cutoff percentage of 98, a static modification on cysteine (+57 Da), and an N-terminal dynamic modification of +43 Da (carbamylation). For the directed searches, MyriMatch was configured to consider a dynamic modification corresponding to an oxidation (+16 Da) on either a methionine or alanine.

For Chapter 6, the TagRecon search algorithm was employed for a peptide sequence tagging method. The TagRecon algorithm requires three types of inputs: MS output files (Thermo .RAW), a protein sequence database, and inferred sequence tags from the output files of the DirecTag algorithm[57]. The DirecTag software uses three major steps: spectrum preprocessing, tag enumeration, and scoring. For preprocessing, the software simplifies each MS/MS spectrum by consolidating isotopic fragment ion packets. Next, DirecTag retains a user-defined number of peaks (100 peaks by default) for each spectrum, filtering out those of low-intensity. Using the remaining peaks, the software identifies pairs of peaks that are separated by amino acid masses. As such, each spectrum represents a graph, where the peaks are represented by nodes and amino acid gaps by edges. To infer sequence tags, a set of nodes that are joined by consecutive edges constitutes a tag. Once all possible tags (of a user-defined length) have been enumerated,

each tag is subjected to scoring. For each tag, the software creates three probabilistic scores: an intensity-based score, mzFidelity score, and a complimentary score. First, the peaks that constitute a tag are evaluated on the basis of their peak intensities. Rather than using an MVH score, DirecTag evaluates the intensity by computed a rank sum. For example, if a tag includes the 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} most intense peaks, the rank sum is 10. This metric is then converted to a *p*-value. Next, similar to SEQUEST and MyriMatch, mzFidelity of the *m/z* spacing is computed for each tag. Since complementary b- and y-fragment ions ($y_i = L+1-b_i$; where *i* is the position of the b-ion, and *L* is the length of the peptide) increase the confidence in peak matching, DirecTag assesses the number and agreement between complimentary ions and then uses information to calculate a *p*-value. In the last step, DirecTag employs Fisher's method to combine all three *p*-values and then uses this joint value to rank all of the tags for each spectrum.

With a sequence tag identified for each MS/MS spectrum, TagRecon processes the information and performs two steps: searching and scoring. Similar to other database search algorithms, TagRecon compares the experimental spectra to virtual spectra created by a user-supplied protein database. Unlike other searching algorithms, a small sequence tag is already available to cull potential candidate sequences from the database. When TagRecon detects that a small sequence tag matches a peptide sequence, the software compares the flanking regions of both the experimental and theoretical spectrum to determine whether the m/z values match within a user-defined mass error. If either of the flanking regions matches, then the candidate sequence may be used to explain the remaining spectrum. Importantly, the algorithm can be parameterized to allow for one mismatch to occur during the mass matching step. If a mismatch occurs, the software computes the difference and explains away the mass shift as either an unexpected mutation or post-translational modification. Similar to MyriMatch, for each MS/MS spectrum, TagRecon computes three scores: MVH, mzFidelity, and XCorr.

For the peptide sequence tagging searches in Chapter 6, DirecTag generated partial sequence tags (tag length = 3) from MS/MS spectra for each output file (Thermo .RAW). DirecTag searches were configured with the following parameters: precursor ion mass tolerance of 0.01 *m/z* units, fragment ion mass tolerance of 0.5 *m/z* units,

complementary ion mass tolerance of 0.5 *m/z* units, a total ion current cutoff percentage of 98, a max tag score of 20, and a max tag count of 50. For each spectrum, TagRecon reconciled the inferred sequence tags against a subset protein database (i.e., proteins identified by the MyriMatch database search) while making allowances for one unanticipated mass shift. TagRecon was configured (mutation mode) to consider mass shifts corresponding to amino acid substitutions using the BLOSUM62 matrix. In addition to mutations, TagRecon was configured to consider the following modifications: infinity tryptic cleavages allowed, precursor ion mass tolerance of 0.01 *m/z* units, fragment ion mass tolerance of 0.5 *m/z* units, a total ion current cutoff percentage of 98, a static modification on cysteine (+57 Da) and an N-terminal dynamic modification of +43 Da.

2.3.2 Protein Inference

After SEQUEST, MyriMatch, and TagRecon database searches were performed, the confident peptide identifications needed to be filtered and assembled into the context of protein identifications. For SEQUEST searches, the DTASelect software[163] was used to apply multiple layers of filtering to the search results and to assemble proteins. In general, the software functions in three phases: summarization, evaluation, and reporting. The first step is to extract and summarize the most important information (i.e., scores and peptide positions in proteins) from each SEQUEST peptide spectrum match. Next, identifications are evaluated by applying used-defined criteria to each peptide spectrum match. If a match passes all the specified criteria, then that peptide sequence is processed at a higher level. Once all confident peptide sequences have been acquired and mapped to their respective proteins, only proteins with a user-specified number of peptides are retained. In the final step, a DTASelect file is created to report all of the proteins, peptides, and PSMs that passed all of the specified criteria. For Chapters 3, 4 and 5, resulting peptide identifications from SEQUEST were filtered and organized into protein identifications using DTASelect version 1.9. Each peptide identification required Xcorr values of at least 1.8 (+1 charge state), 2.5 (+2 charge state), or 3.5 (+3 charge state) and a DeltCn $\geq$0.08, and only proteins with two peptides sequences were retained.

For MyriMatch and TagRecon searches, the IDPicker software[164] was employed to refine search results and assemble peptides into proteins. Unlike DTASelect, IDPicker was not built to rely on peptide identification score thresholds. Instead, the software relies on a PSM- and protein-level FDR to control the quality of the peptide and protein identifications, and thus employs a dynamic threshold. The latest version of IDPicker (v3) incorporated 3 modules: PSM-level FDR calculation (IDPqonvert), protein assembly and filtering (IDPassemble), and reporting (IDPreportFDR). In the first step, the software extracts peptide, sequence, scan, and scoring information from MyriMatch output files. Next, the software determines identification score thresholds that correspond to a user-specified FDR cutoff. In the protein assembly step, the IDPicker assembles peptides into proteins and applies user-defined protein-level filters – this includes minimum spectra per peptide, minimum spectra per match, maximum protein groups, maximum distinct peptides, minimum additional peptides, and minimum spectra per protein. In the context of IDPicker, a "distinct peptide" is a peptide that is not only unique to the database, but that also has a unique mass. In other words, charge states and modifications to unique sequences increase the number of distinct peptides. In addition, the "additional peptides" criteria can be applied to enforce parsimony – that is, IDPicker will create a list of protein groups, which provides a minimal list of proteins that sufficiently explain all of the identified peptides. For Chapter 6, resulting peptide identifications from MyriMatch and TagRecon were filtered and organized into protein identifications using IDPicker version 3. IDPicker filtered the resulting peptide-spectrum matches (PSMs) from all searches at a 2% FDR. While search algorithms rigorously assess the statistical significance of each PSM, high-throughput validation of modified peptides remains an open problem. In this study, we applied tested attestation principles for validating modified peptides in a complex mixture[165]. To obtain a data set of the highest quality, we enforced the following filtering guidelines: (1) mutated peptides were removed if they mapped to a contaminant protein, (2) mutations of lysine or arginine residues cannot occur at trypsin cut sites, (3) if a spectrum matched to a mutated peptide (TagRecon) as well as a non-mutant (MyriMatch) peptide, the mutated PSM must improve upon the score of the unmodified PSM by 10%, (4) a distinct mutated peptide sequence must match to at least 3 different

spectra, and (5) mutations that can also be explained as common sampling processing artifacts were removed: these included the deamidation (+1 Da) of asparagine or glutamine, dehydration (-18 Da) of aspartate and glutamate, formylation (+28 Da) of threonine or serine, and the oxidation of methionine (+16 Da). Only peptides passing the FDR threshold and the above guidelines were considered for further analysis. Lastly, only protein identifications with at least two distinct peptide identifications were considered for further analysis.

Once proteins have been identified by at least two peptides, they were categorized by their level of uniqueness to the database using the principles of parsimony with Occam's razor constraints – that is, each protein was classified by its level of ambiguity. As previously discussed, proteins are rationally organized using the following nomenclature: proteins that consist of only distinct peptide identifications were classified as *distinct* proteins. Proteins were classified as *differentiable* when they contain at least one peptide that is unique to that locus, as well as one or more peptides that map elsewhere in the proteome. The *indistinguishable* proteins consisted of measured non-unique peptides that map elsewhere in the data set.

2.3.3 Creation of Protein Groups

Although a protein-level classification system provides a highly accurate and readily interpretable protein summary report for microbial proteomes, strict implementation of the Occam's razor approach can be misleading when applied to higher eukaryotic organisms. The current limitation of this approach for eukaryotic organisms is that the number of confident protein identifications (i.e., proteins with at least one distinct peptide) is severely reduced because of the prevalence of shared peptides. As a property of evolution, genetic redundancy is rampant across the eukaryotic kingdom. In fact, many eukaryotic organism genomes have been duplicated more than once in their evolutionary past. As a result, the majority of genetic redundancy observed is between gene homologues. Immediately after gene duplication, these genes (i.e., proteins) are believed to be functionally redundant. It is generally assumed that one of the redundant genes is initially free of all selective pressure, allowing the gene to acquire advantageous (i.e.,

neofunctionalization) that may lead to a new function. Consequently, the function of some duplicated genes may only be partially redundant.

Since genes (i.e., proteins) with extensive sequence similarity have a high likelihood of performing similar biological roles in a cell, they can be collapsed together by sequence homology algorithms. By grouping homologous proteins together, this consolidates indistinguishable proteins into a meaningful report, while preserving biological information. The research presented in this dissertation has demonstrated that this provides a means to alleviate the majority of ambiguity associated with shared peptides. Similar to a peptide being unique to a protein within the database, many shared peptides are found to be unique to a particular protein group. Although in some cases it may not be clear as to which member of a protein group is actually present in a given sample, the identification of peptides belonging to a particular protein group likely indicates the presence of a shared functional process, especially considering the relatively stringent similarity cut-off (90%) that's commonly applied. Despite sacrificing some level of protein resolution, this approach accurately resolves protein ambiguity as a result genetic redundancy

For Chapters 3 through 6, *P. trichocarpa* database proteins sharing extensive sequence homology were assigned to protein groups using a freely-available software package, USEARCH v5.0[166]. Proteins that shared more than 90% of their sequence with another protein in the database were clustered by pairwise sequence comparisons using the UCLUST program (similarity threshold of 0.9). A similarity threshold of 0.9 was chosen to reflect the level of intraproteomic similarity in the *Populus* proteome: two genome-wide duplication events have increased the level of redundancy, in which nearly two-thirds of the protein-coding genes share sequence similarity (>90%)[112]. Protein groups are defined by the longest protein sequence, the seed, which shares ≥90% sequence identity to each protein in that cluster. All groups were manually verified to ensure that obvious redundancies, such as alternatively spliced variants, remained together. Identified peptides that were distinct to a particular protein group were marked as protein-group unique, meaning these peptides did not belong to another group of proteins. All peptides that were originally database-unique were necessarily protein

group-unique, but grouping peptides from homologous proteins allowed shared peptides to be considered distinct if they belonged to only one protein group identification (i.e., a protein-group unique peptide.

2.3.4 Protein Quantitation

For semi-relative quantitation measurements, protein abundances were measured by spectral counting. As described previously (*vide supra*), spectra counts (SpC) are defined as the number of times a peptide was observed in a given chromatographic experiment. Therefore, the number of observations of a protein's constituent peptides can be used to compare a protein's relative abundance between two samples[167]. Any increase/decrease in that protein's summed SpC between samples can be interpreted as a relative change in abundance; however, changes in relative protein abundances can also stem from systematic errors in experimentation and/or instrumentation. Ideally, when comparing protein abundance differences between two proteomes, every protein and peptide should experience the same variability. While certain steps are taken to minimize variability, the physiochemical properties that make proteins and peptides so functionally different are also the source of experimental biases. Clearly, proteins can vary in length and amino acid composition and, as such, they will likely generate not only a different number of peptides, but also a different set of peptides for sequencing. Therefore, spectral counting is only a powerful method when comparing protein A in *proteome 1* and *proteome 2*. In this scenario, it's safe to assume that this protein will have the same experimental biases (protein solubility, enzymatic digestion potential, etc.) as well as instrumental biases (ionization efficiency, retention time, fragmentation patterns, etc.) in the two conditions. As a result, if peptides for a particular protein decrease in SpC abundance between conditions, it is acceptable to assume the protein has decreased in abundance[168]. It is also important to stress that one cannot assume that a protein's low SpC always reflects low abundance in the sample; its peptides may just be poorly retained by the LC separation or not MS compatible.

Although variability cannot be completely eliminated, it can be measured, and reduced to a certain degree. By maintaining proper quality control during sample preparation and MS analysis, only a small degree of variation should be expected. To

identify and gauge the degree of biological or technical variation, it is best to include biological replicates and technical replicates for each analysis. Since there are recognizable sources of variability, normalization approaches are commonly applied to improve the agreement between peptide ratios observed across instrument runs. The most common method converts SpC values for each protein into a normalized spectra abundance factors (NSAFs)[169]. In brief, this method is based on the total spectral counts reported (normalize for run to run variation) and the size of a protein (normalizes by protein length because larger proteins contribute more peptides/spectra). Therefore, the NSAF is calculated as the number of SpCs identifying a protein, divided by the protein's length, divided by the sum of SpC/L for all proteins in the experiment.

For this dissertation, NSAF values were calculated for each protein group identified by normalizing the sum of the total spectral counts for each peptide belonging to a protein group. For peptides belonging to multiple protein groups, the spectral counts were recalculated based on the proportion of uniquely identified peptides between the protein groups sharing the peptide in question. While NSAF values typically account for biases resulting from protein length, here NSAF values were calculated for each protein group by using the length of the seed sequence. Adjusted NSAF (nSpC) values were calculated for each protein group[86, 170].

CHAPTER 3

DEFINING THE BOUNDARIES OF FUNCTIONAL GENOME EXPRESSION IN POPULUS USING BOTTOM-UP PROTEOMICS

All of the data presented below has been adapted from the following published journal article:

Paul Abraham, Rachel Adams, Richard Giannone, Udaya Kalluri, Priya Ranjan, Brian Erickson, Manesh Shah, Gerald Tuskan, Robert Hettich. "Defining the Boundaries and Characterizing the Landscape of Functional Genome Expression in Vascular Tissues of Populus using Shotgun Proteomics". *Journal of Proteome Research* ***2012*** 11(1): 449-460. Sample preparation and mass spectrometry experiments were performed by Paul Abraham. The bioinformatic workflow for protein grouping was developed by Paul Abraham, Rachel Adams, and Richard Giannone. The in-house scripts for protein grouping were created by Rachel Adams. Biological data analysis was performed by Paul Abraham.

3.1 Introduction to *Populus* Proteomics

The advent of high-throughput DNA sequencing has revolutionized the assembly of high-quality genomes for prokaryotes and eukaryotes such as plants and humans[171]. The release of reference genomes (http://www.phytozome.net) has paved the way for "omics"-based research, which has focused on the identities and functions of the suites of genes and proteins that are important for plant growth and development[172]. In particular, the rapidly developing field of proteomics is already providing remarkable insight into cellular activities at the protein level that complement genomic and transcriptomic investigations[173-175]. That is, obtaining deep protein-level measurements for the identification, quantification, post-translational modification, and localization of proteins has facilitated a more comprehensive understanding of molecular functionality. While there are a variety of proteomic techniques available to measure protein abundance, they differ greatly in their analytical merits of sensitivity, depth of measurement, resolution, and throughput.

Following the release of the *Populus trichocarpa* genome in 2006, *Populus* emerged as a model system for the study of woody perennial plant biology[176]. The availability of a sequenced genome has prompted vigorous proteomic investigations

aimed at elucidating developmental phenomena pertinent to *Populus*[125, 177-179]. Here, we investigate the growth and development of the tree vascular network, which involves a complex system that integrates both molecular signaling components and regulation of protein expression. In higher plants, this elaborate network exists in two vascular tissues, phloem and xylem. Spanning the entire length of plants, these extensive vascular networks are responsible for the distribution of water and essential nutrients across long distances to vital locations. Insights derived from the detailed identification of proteins and their abundances within *Populus* vascular tissues will undoubtedly yield an improved understanding of the growth and development processes, such as wood biogenesis and drought response.

The full potential of shotgun proteomics in plants is limited in part by the complexities of the proteomic reference database. Most plant genomes contain functional gene redundancies, segmental duplications, whole-genome reorganizations, and single nucleotide polymorphisms (SNPs) that have led to adaptive specialization of pre-existing genes (i.e., gene models, protein families and gene duplications that share >90% sequence identity). This inherent redundancy within all plant proteomes confounds the accuracy of the proteome characterization, inflating the total number of proteins identified and/or leading to incorrect biological interpretations. A sophisticated bioinformatics workflow for assigning peptides to proteins and for interpreting resulting protein identifications has to be employed to deal with gene duplications and extended gene families in *Populus*.

Database searching algorithms, such as SEQUEST and MASCOT, which are commonly used to match experimental tandem mass spectra to theoretical fragmentation spectra generated from a pre-defined proteomic sequence database, cannot resolve peptide spectral matching for any peptide variation unaccounted for in the database. Therefore, when dealing with higher eukaryotes such as humans[180] and plants, a major issue for tandem MS and peptide identification algorithms is the high level of sequence variation, including naturally occurring post-translational modifications (PTMs) and SNP-based single amino acid polymorphisms (SAAPs). In many proteomic measurements, such as those for microbial species, modifications and peptide isoforms

do not dramatically affect proteome identification and thus are ignored. In contrast, the complexities of plant proteomics demand attention to these protein alterations, as they have a significant impact on the quality of the proteome characterization[181]. Thus, the degree of sequence variation in *Populus* was explored to identify a number of unassigned quality spectra that result from these common peptide modifications.

In this study, current experimental and computational approaches were employed to obtain a broad proteome profile of *Populus* vascular tissue. The experimental context includes 1) a large *Populus* sample set consisting of two genotypes grown under normal and tension stress conditions[182], 2) bioinformatics clustering to effectively handle gene duplication, and 3) an informatics approach to track and identify single amino acid polymorphisms. Together, the integration of deep proteome measurement on an extensive sample set with protein clustering and characterization of peptide sequence variants has provided a level of proteome characterization for *Populus* that has not yet been observed.

3.2 Characterizing the Landscape: Global Survey of the *Populus* Proteome

3.2.1 Mapping Deep Measurements to the Populus Proteome

To generate a high-coverage proteome profile, we performed bottom-up proteomics on a large sample set consisting of subcellular fractions (soluble, pellet) of two tissue types (xylem, phloem) from two *Populus* species: *P. deltoides and P. tremula x alba*. Using the most recent *Populus* genome draft (v2.0, http://www.phytozome.net/cgi-bin/gbrowse/poplar/), tandem mass spectra from 60 *Populus* proteome measurements collectively identified 7,505 total proteins and 33,233 tryptic peptide sequences with an overall false discovery rate of <1% at the protein level. Combining the proteome measurements together provided a global view of protein expression involved in vascular tissue development, resulting in protein assignments for ~17% of the predicted *Populus* proteome. Approximately 40% of all detected proteins belonged to three specific functional categories based on 24 EuKaryotic Orthologous Groups (KOGs): 1) unknown function, 2) post-translational modification and turnover, and 3) signal transduction (Figure 3.1). The remaining identified proteins are scattered

across the other functional categories. The number of redundant proteins and peptides identified for each sample type and technical replicate are shown in Table 3.1.

3.2.2 Genetic Redundancy and Protein Classification

Shotgun proteomics employs a peptide-centric approach that relies on the ability to accurately assemble and assign thousands of measured peptides to reference proteins in biological samples. Although this is the conventional method for identifying proteins in large-scale studies, this approach presents several challenges when assigning peptides to proteins in higher eukaryotes. The most common issue deals with inferring a protein's existence through the identification of peptides that constitute its primary structure. Protein inference becomes problematic when two or more proteins share peptides[72-73, 183]. Shared or degenerate peptides are natural occurrences that originate from protein homology, conserved protein domains among various proteins, splice variants, and redundant entries due to gene duplication events, all of which are common in plants[184-185]. Compared to *A. thaliana*, the *Populus* genome is highly genetically redundant, such that two-thirds of protein-coding genes share sequence similarity greater than 90% (Figure 3.2A-B). After performing an *in silico* digest of the *A. thaliana* protein reference database, there were ~4.3 million fully tryptic peptides in the database. Out of those, ~320,000 peptides are shared between two or more proteins. After completing an in silico digest of the *P. trichocarpa* reference protein database, ~6.3 million fully tryptic peptides were present and, of those, ~2 million are shared between two or more proteins.

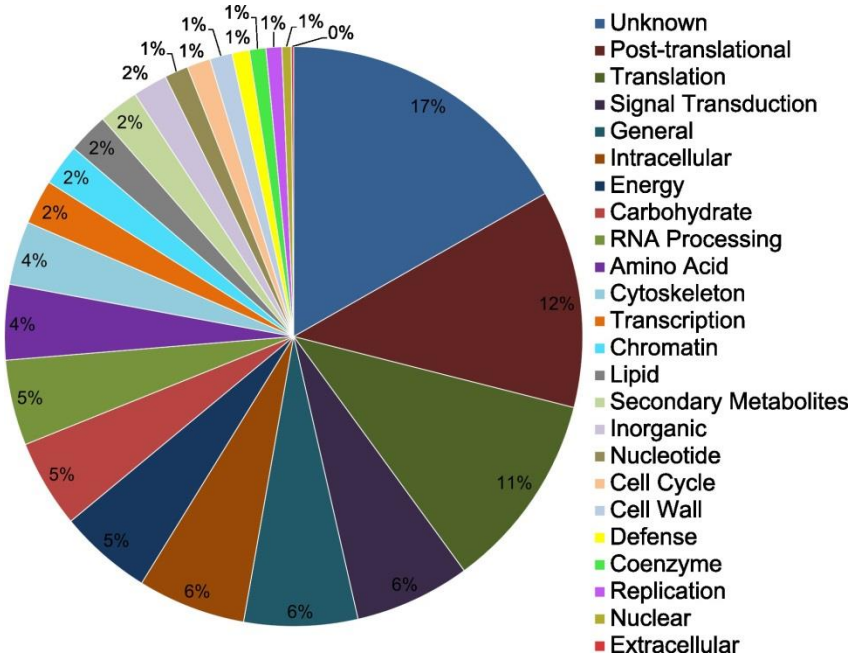


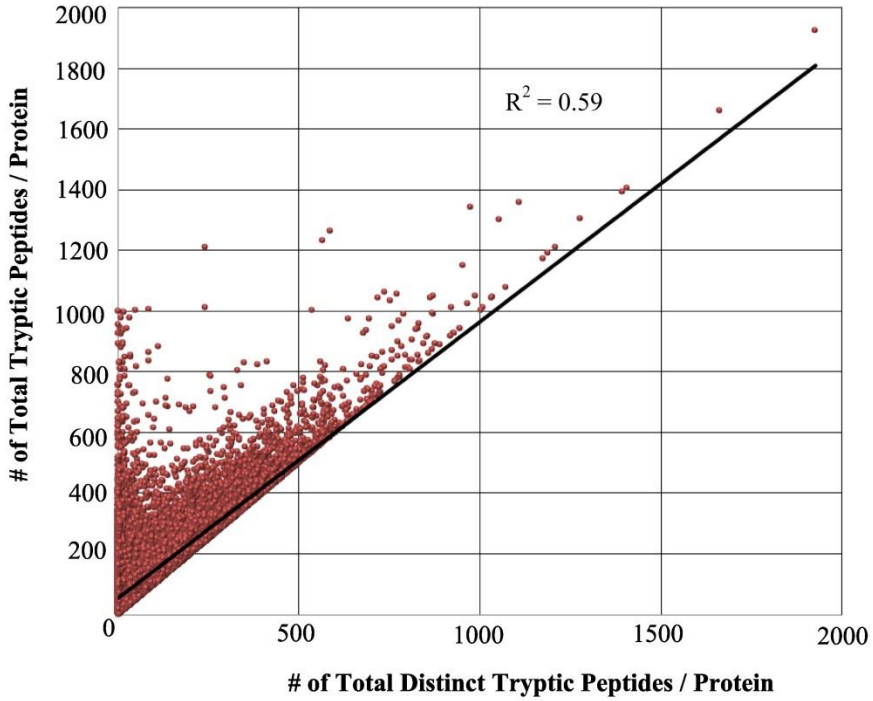
Figure 3.1. Distribution of detected proteins by their KOG functional classification categories. The data indicates the most abundant functional categories for the combined xylem and phloem vascular tissue proteomes.

Table 3.1. The total number of proteins and peptides observed for each of the twelve conditions in two genotypes

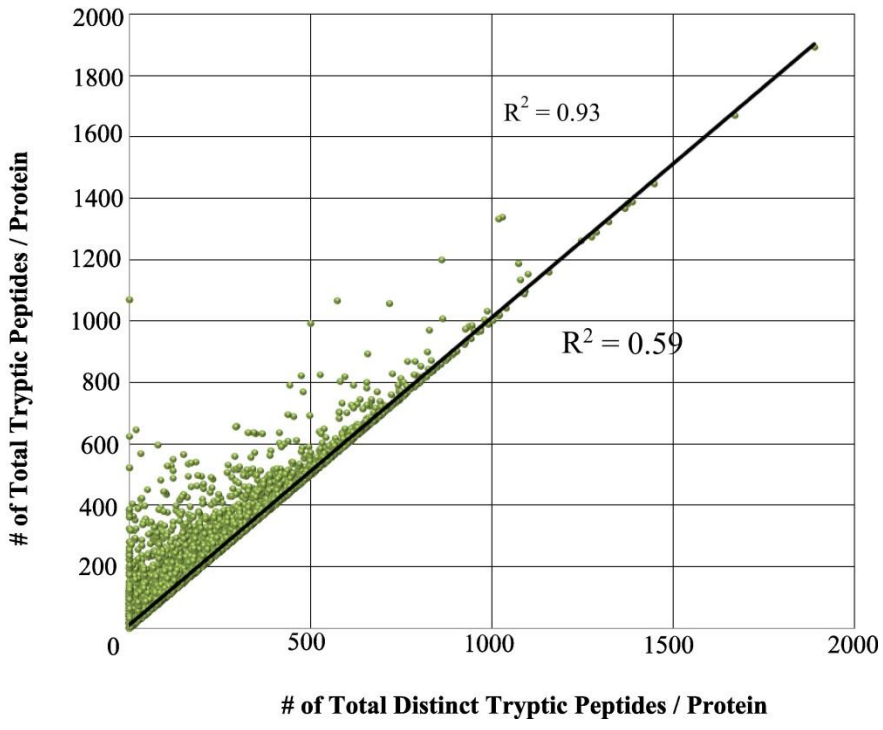
	P. deltoides		*P. tremula X alba*	
Sample Type	Proteins	Peptides	Proteins	Peptides
Xylem Stress (Normal) Soluble Replicate 1	3690	18715	2980	14694
Xylem Stress (Normal) Soluble Replicate 2	3623	18371	2795	12079
Xylem Stress (Tension) Soluble Replicate 1	3088	17278	2889	13497
Xylem Stress (Tension) Soluble Replicate 2	3200	17730	3048	17223
Xylem Stress (Opposite) Soluble Replicate 1	3846	20579	3032	16929
Xylem Stress (Opposite) Soluble Replicate 2	3608	19541	2106	9683
Xylem Stress (Normal) Pellet Replicate 1	2145	8353	1631	6584
Xylem Stress (Normal) Pellet Replicate 2	2304	9491	1467	5549
Xylem Stress (Normal) Pellet Replicate 3	2325	9418	894	2654
Xylem Stress (Tension) Pellet Replicate 1	2267	10923	1714	5491
Xylem Stress (Tension) Pellet Replicate 2	2092	9618	1949	7843
Xylem Stress (Tension) Pellet Replicate 3	2130	9735	1477	4724
Xylem Stress (Opposite) Pellet Replicate 1	2340	9902	1380	4250
Xylem Stress (Opposite) Pellet Replicate 2	2276	9582	1680	5641
Xylem Stress (Opposite) Pellet Replicate 3	2203	9627	1835	6680
Phloem Stress (Normal) Soluble Replicate 1	2322	8401	1676	6198
Phloem Stress (Normal) Soluble Replicate 2	2237	8314	1243	4021
Phloem Stress (Tension) Soluble Replicate 1	2798	10995	1585	5538
Phloem Stress (Tension) Soluble Replicate 2	2840	11519	1412	4557
Phloem Stress (Opposite) Soluble Replicate 1	2396	8875	1328	4270
Phloem Stress (Opposite) Soluble Replicate 2	2432	9094	1657	6128
Phloem Stress (Normal) Pellet Replicate 1	2038	5706	1895	5808
Phloem Stress (Normal) Pellet Replicate 2	2091	5964	788	1591
Phloem Stress (Normal) Pellet Replicate 3	1944	5699	1335	3327
Phloem Stress (Tension) Pellet Replicate 1	2314	7176	529	903
Phloem Stress (Tension) Pellet Replicate 2	2158	6500	347	503
Phloem Stress (Tension) Pellet Replicate 3	2164	5636	500	823
Phloem Stress (Opposite) Pellet Replicate 1	2124	6860	1824	5450
Phloem Stress (Opposite) Pellet Replicate 2	2054	6651	1864	5123
Phloem Stress (Opposite) Pellet Replicate 3	1563	4149	2027	5898

Figure 3.2.Illustration of the degree of intraproteomic similarity for (A) *P. trichocarpa* (red) and (B) *A. thaliana* (green). Each circle represents a protein in the organism's reference protein database. The (C) distinct and differentiable protein identifications (3,510) and the (D) indistinguishable protein identifications (3,995 were mapped onto the graph.

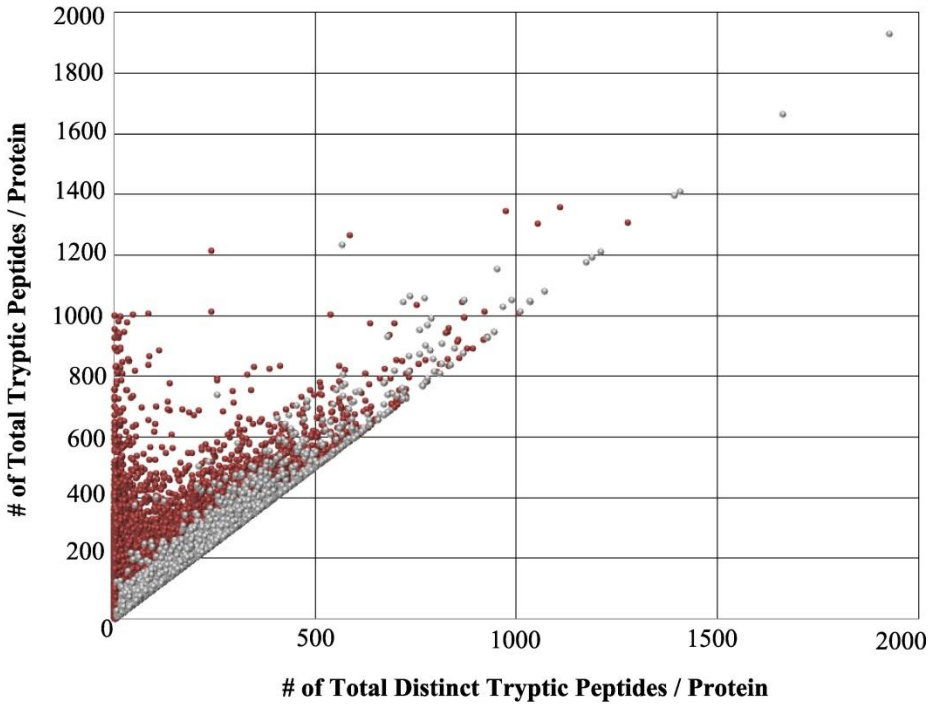
(A)


2000
1800
1600
1400
1200
1000
800
600
400
200
0
500
1000
1500
2000
$R^2 = 0.59$
# of Total Tryptic Peptides / Protein
# of Total Distinct Tryptic Peptides / Protein

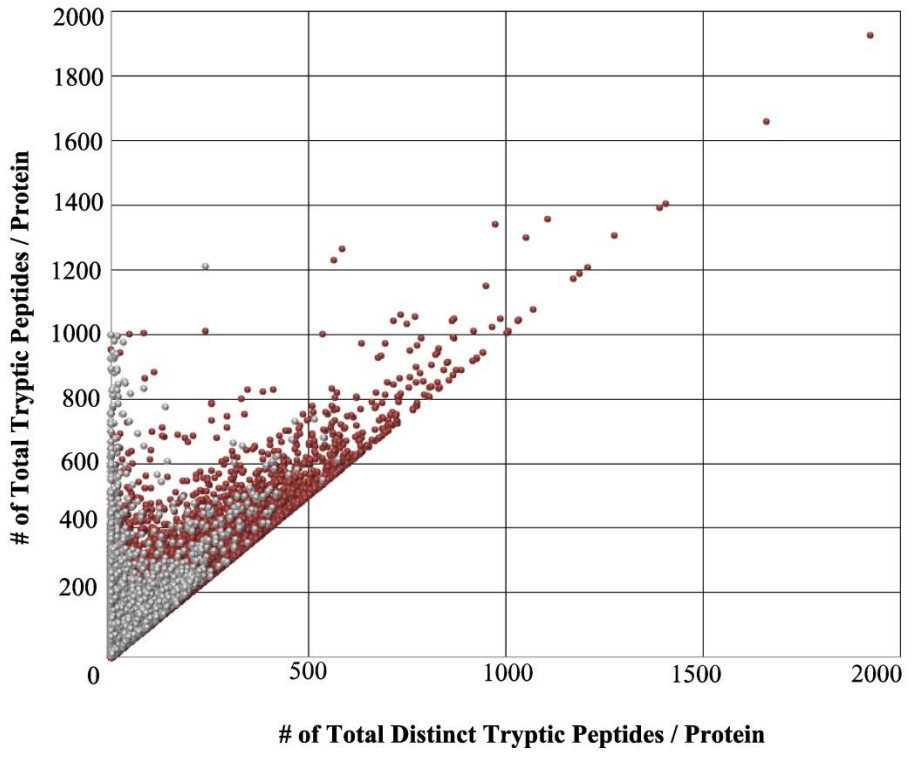
(B)


2000
1800
1600
1400
1200
1000
800
600
400
200
0
500
1000
1500
2000
$R^2 = 0.93$
$R^2 = 0.59$
# of Total Tryptic Peptides / Protein
# of Total Distinct Tryptic Peptides / Protein

(C)


2000
1800
1600
1400
1200
1000
800
600
400
200
0
500
1000
1500
2000
# of Total Tryptic Peptides / Protein
# of Total Distinct Tryptic Peptides / Protein

(D)


2000
1800
1600
1400
1200
1000
800
600
400
200
0
500
1000
1500
2000
# of Total Tryptic Peptides / Protein
# of Total Distinct Tryptic Peptides / Protein

Clearly, the level of genetic sequence redundancy is extensive in the *Populus* proteome. Therefore, within these large data sets emphasis must be placed on accurate identification and validation of proteins, accounting for highly conserved, shared peptides.

In previous studies, the categorical nomenclature of Yang et al. (2004) has been adapted to rationally organize the peptide data from each LC-MS/MS experiment. Several research groups have shown that this nomenclature can be coupled with Occam's razor constraints to provide a minimal list of proteins to explain all observed peptides[73]. Using this classification method, we consolidated protein assignments by their level of uniqueness. Proteins that consist of only uniquely identified peptides were classified as distinct proteins. Proteins were classified as differentiable when they contain at least one peptide that is unique to that locus, as well as one or more peptides that map elsewhere in the proteome. The indistinguishable proteins consisted of only measured non-unique peptides that map elsewhere in the data set. Within our entire data set, only 50% of the tryptic peptides identified were classified as unique to the database. Therefore, out of the 7,505 total protein identifications in the present study, 3,510 proteins were uniquely identified (classified as distinct or differentiable) and 3,995 proteins were categorized as non-unique or indistinguishable (Figure: 3.2C-D).

Using the nomenclature above, we generated a minimal list of proteins that were conclusively determined to be present within the data set. However, due to the inherent ambiguity of the *Populus* proteome, less than 50% of the proteins categorized by the above-mentioned criteria could be used for biological interpretation. In addition, due to the extensive homology within the database, a vast majority of the proteins were classified as indistinguishable. As most of the proteins in this category contain no unique peptides, it was difficult to determine which specific proteins were present in the sample using an MS-based approach. As shown in other studies, one approach for proteins that cannot be distinguished on the basis of identified peptides is to collapse these into protein groups to provide a more accurate and informative data set[186-187]. In an attempt to reconcile this problem, a bioinformatics workflow was incorporated to better handle proteins sharing high sequence homology (≥90%) to increase qualitative accuracy by avoiding the over- and under-identification of homologous proteins.

An illustration of the informatics workflow can be seen in Figure 3.3A. Briefly, proteins sharing 90% or more sequence identity were clustered into groups by UCLUST, a clustering algorithm functionally equivalent to BLASTP[188]. Each protein group was defined by a representative protein sequence called a seed, where each seed shares ≥90% sequence identity to each protein in that cluster. By applying the clustering algorithm to the *Populus* database, the number of protein entries decreased from 64,689 proteins to a total of 43,069 protein groups. Implementation of clustering to the data set reduced the 7,505 observed proteins to a total of 4,226 protein groups (see Methods), in which 2,016 were singletons (i.e., a one-member group). This reduction implies that ~50% of the observed proteins were clustered into groups that shared extensive sequence homology. Therefore, this approach effectively consolidates indistinguishable proteins into a meaningful report. Although grouping proteins by high sequence similarity undoubtedly sacrifices some level of protein resolution, it is reasonable to assume that proteins with this level of sequence homology share similar biological functions. Furthermore, integrating the clustering approach with the initial SEQUEST analysis provided a means to categorize which members of a protein group were unique.

Due to the peptide-centric nature of shotgun proteomics, it was imperative to report peptides in the context of proteins groups. As expected, clustering proteins into groups alleviated some of the ambiguity associated with shared peptides. Similar to a peptide being unique to a protein within the database, we found many peptides were unique to a particular protein group within the clustered database. In fact, 68% of previously shared peptides that were classified as non-unique to the Populus database were reclassified as unique to the clustered database. Moreover, the bioinformatics workflow generated a data set where 84% of the detected peptides were classified as unique. Therefore, rather than disregarding these peptides from the analysis, they were rescued and used for biological insight (Figure 3.3B). While it may not be clear as to which member of a protein group is actually present in a given sample, the identification of peptides belonging to a particular protein group likely indicates the presence of a shared functional process, especially considering the relatively stringent similarity cut-off (90%) applied to the protein database[189].

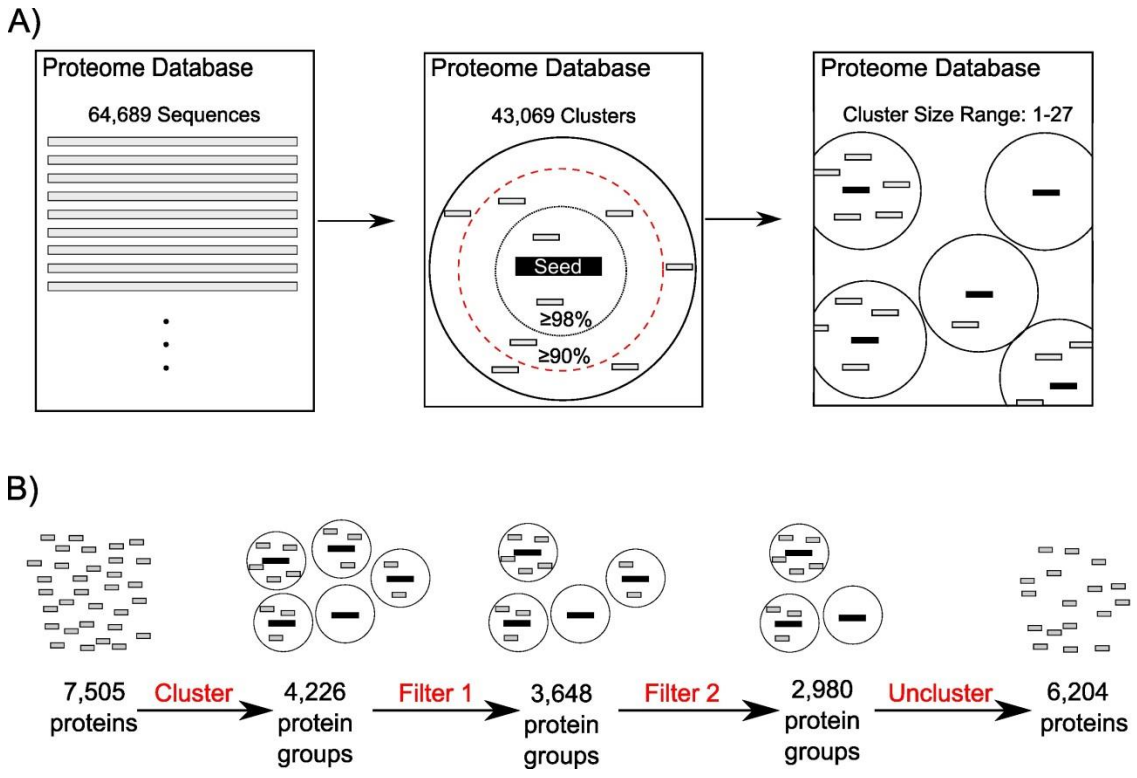


Figure 3.3. Illustration of the protein grouping bioinformatic workflow. A) All proteins in the proteomic database were clustered by UCLUST to deal with gene duplications and extended gene families. B) After the proteins were clustered into protein groups, a conservative two-tiered filtering approach was used to eliminate 1) ambiguous identifications and 2) those at the lower detection limits.

3.2.3Characterization of the Populus Vascular Tissue Proteome

Xylem and phloem tissues are responsible for long-distance transportation and storage of essential minerals and nutrients in plants. A recent study used bottom-up proteomics to examine proteins expressed during xylem development[128]. This approach demonstrated an ability to robustly characterize xylem tissue in *Populus* by vastly increasing the number of proteins identified and characterized relative to previous *Populus* proteome studies[125]. In the current study, a similar experimental approach was applied to identify and contrast the relationship and dissimilarities between the xylem and phloem proteomes. A "core" proteome was extracted from the entire data set, consisting of 2,627 protein groups that were confidently identified in both xylem and phloem. The core proteome, encompassing 59% of the total proteins identified in the Populus data set, includes proteins representing each KOG category (Figure 3.4). The core metabolic signature is consistent with other studies that show an overrepresentation of proteins that are involved in energy production and translation[79]. Moreover, a similar quantitative distribution profile was also observed during xylem development[128, 190]. In addition, these functionally and spatially separate vascular networks contain tissue-specific proteins: 606 unique xylem proteins-groups and 461 unique phloem protein groups, each having a distinct metabolic profile as shown in Figure 3.4.

3.2.4 Regulatory Proteins Involved in Vascular Tissue Development

Among the proteins identified in the *Populus* data set were proteins that have been shown to control the patterning and differentiation of vascular tissues. Interestingly, the receptor protein kinase CLAVATA1 precursor, a part of the CLV3/CLV1 system, was exclusively identified in phloem tissue. The developmental process of the plant vascular network is a complex system that integrates both molecular signaling components and regulation of protein expression. Stem cells in the shoot apical meristem regulate the continuous formation of the different tissues during vascular formation. It has been shown that the receptor protein kinase CLAVATA1 governs stem cell fates in the shoot apical meristem. Along the boundaries of the procambium/cambium space of postembryonic tissue, this process occurs when CLAVATA1 binds to the protein ligand CLE41, which is secreted from the phloem[191-192].

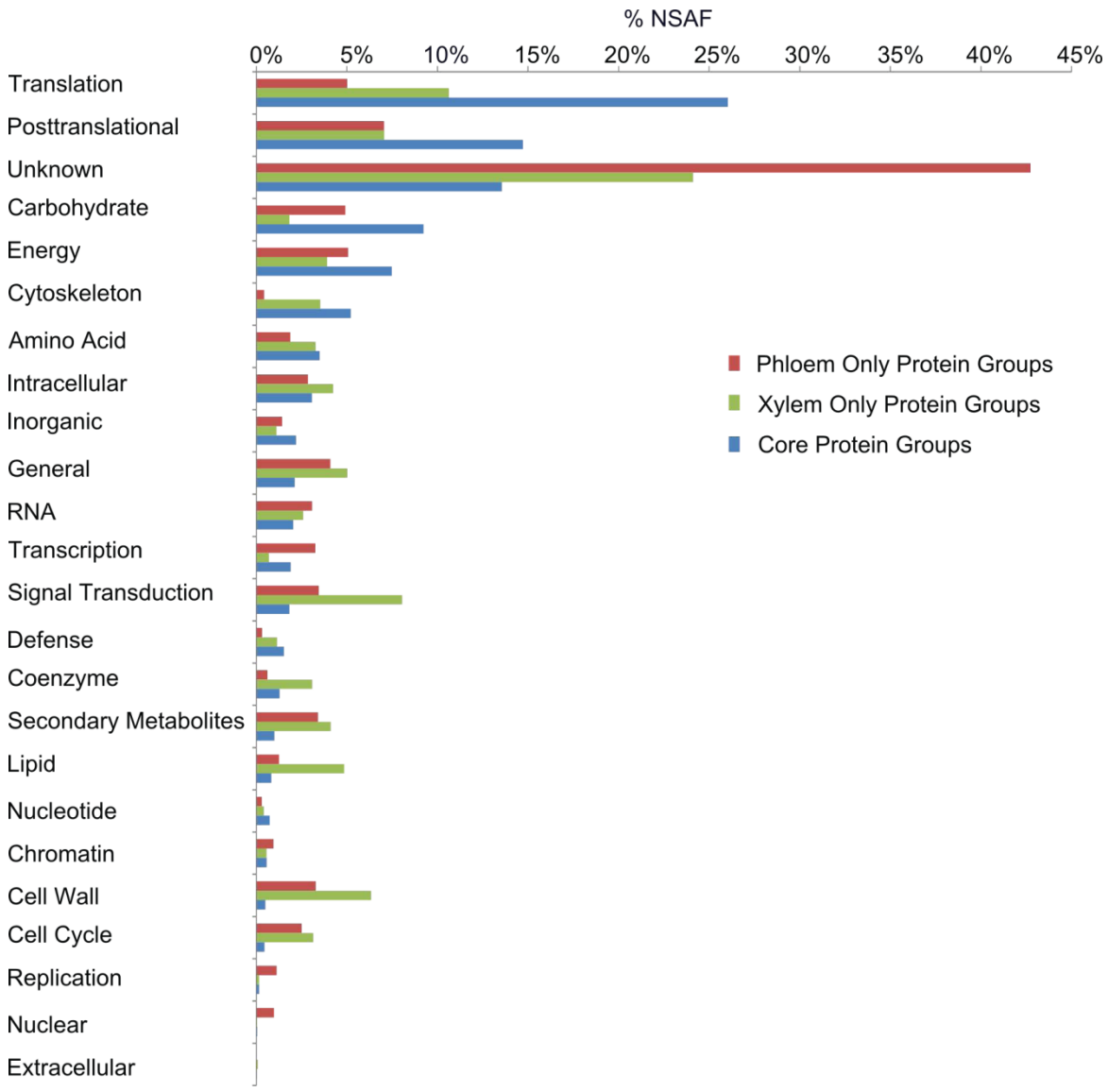


Figure 3.4. Quantitative distribution of detected proteins by their KOG functional classification category. The relative abundance of each functional category was calculated as a percent of the summed protein group abundance within each classification: protein groups found in xylem and phloem (the core proteome), protein groups found only in phloem, and protein groups found only in xylem.

We also identified bri1 suppressor 1 (BSU1) protein only in xylem tissue. BSU1 is a ser/thr-protein phosphatase that has been shown to be a positive regulator of brassinolide signaling, thereby playing an important role in the regulation by brassinosteroids. It has been shown that brassinosteroids regulate xylem differentiation and vascular patterning from cambium cells[193]. Furthermore, brassinosteroid lack-of-function mutants in *Arabidopsis*[194] and rice[195] disrupt vascular development. It is also known that the plant hormone auxin plays a critical role in the cell-to-cell communication in vascular differentiation[196]. We detected evidence for peptidyl-prolyl cis-trans isomerase protein (PIN1) expression in both xylem and phloem tissue. Currently, many studies suggest that the formation of plant vascular networks is an auxin-transport-based mechanism and the driving force behind this mechanism is the accumulation and polarization of PIN1, an auxin efflux carrier[197-198]. Based on our measurements, PIN1 expression in phloem may provide a bi-directional pathway for long-distance transportation, while expression in xylem leads to vascular development and xylem differentiation.

3.2.5 Biosynthesis and Development of Wood Cell Walls

We identified several of the cell wall-related carbohydrate active enzymes within our data set, including cellulose synthases, pectin methyl esterases, and xyloglucan endotransglucosylases and hydrolases. Wood, or secondary xylem, is a water conduit formed from the vascular cambium that provides mechanical support for plants and is the primary source of chemical feedstock for the emerging biofuels industry[199-200]. The cell wall is composed of a carbohydrate matrix consisting of cellulose microfibrils that are embedded within a mixture of hemicellulose and lignin, a polymer with subunits of phenylpropanoid[201-202]. Carbohydrate active enzymes (CAZymes) are known components of the construction and remodeling of the carbohydrate matrix[203]. Our proteomics profile identified several genes encoding CAZymes, concurrent with results from EST and microarray analysis[204-206].

Lignin, the other main constituent of the wood cell wall, is a complex phenolic polymer that provides a physical barrier that protects plants from microbial and physical attack and provides mechanical support. Lignin is polymerized from three primary

monomers: p-coumaryl alcohol (H), coniferyl alcohol (G) and sinapyl alcohol (S). The monolignols are synthesized from phenylalanine through the phenylpropanoid pathway and, within the *Populus* genome, 95 gene models have been identified as putative phenylpropanoid biosynthesis genes[207]. The genetic and biochemical role of most of the 95 gene models remains undefined. Our study identified proteins associated with the monolignol biosynthesis pathway, identifying members for each enzyme family (Table 3.2).

3.3 Defining the Boundaries: Interrogation of Unassigned MS/MS Spectra

3.3.1 Spectral Quality Assessment

Although remarkable depth of coverage of the *Populus* proteome has been achieved, one of the greatest heuristics that contributes to the success of database-searching approaches also has a complementary limitation: regardless of the quality of peptide-derived spectra, algorithms will only match spectra to peptides that exist within user-defined sequence variations. Peptide sequencing by mass spectrometry is most commonly performed via collisional-induced dissociation (CID), in which peptide ions fragment in a predictable manner to produce dissociation products that yield sequence information. Though widely used for its simplicity and effectiveness, more than 50% of MS/MS spectra collected in a typical shotgun proteomic experiment do not result in high-confidence peptide identifications when using automated search algorithms such as SEQUEST or MASCOT. Even though these low identification rates can be partially explained by the presence of spectra arising from concurrent fragmentation of multiple precursor ions, incomplete fragmentation of peptides, and chemical noise, a large fraction of peptide-derived spectra remain unassigned because of the quality and completeness of the proteome database[73, 208].

Neither prokaryotic nor eukaryotic protein databases typically include protein isoforms or alterations/modifications, and furthermore their omission has a more dramatic effect on higher eukaryotes in which sequence variations and unexpected splice variants are more prevalent.

Table 3.2. Protein and peptide classification for the monolignol biosynthesis pathway.

Protein Family	Protein	Protein Parsimony	Seed	DU peptides	PGU peptides	NU peptides	Total peptides
PAL	pt017188m	*Differentiable*	pt002727m	15	12	14	41
	pt002727m	*Differentiable*	pt002727m	19	11	12	42
	pt026599m	*Differentiable*	pt026599m	16	10	12	38
	pt011283m	*Differentiable*	pt026599m	7	36	11	54
	pt011320m	*Differentiable*	pt026599m	6	33	9	48
C4H	pt030573m	*Indistinguishable*	pt009878m	0	31	0	31
	pt009878m	*Differentiable*	pt009878m	10	16	0	26
	pt030574m	*Indistinguishable*	pt009878m	0	31	0	31
4CL	pt017853m	*Distinct*	pt017853m	2	0	0	2
	pt038100m	*Distinct*	pt038100m	27	0	9	36
	pt023497m	*Distinct*	pt023497m	7	0	0	7
	pt024040m	*Differentiable*	pt024040m	4	0	4	8
HCT	pt023671m	*Distinct*	pt023671m	7	0	3	10
	pt038643m	*Differentiable*	pt038643m	13	0	3	16
C3H	pt002886m	*Indistinguishable*	pt002886m	0	2	3	5
	pt002890m	*Indistinguishable*	pt002886m	0	2	3	5
	pt017558m	*Distinct*	pt017558m	13	0	4	17
CCoAOMT	pt005042m	*Differentiable*	pt005042m	8	17	1	26
	pt039874m	*Differentiable*	pt005042m	4	17	1	22
	pt027738m	*Distinct*	pt027738m	7	0	4	11
CCR	pt005074m	*Indistinguishable*	pt004827m	0	2	0	2
	pt004953m	*Indistinguishable*	pt004827m	0	4	0	4
	pt004827m	*Indistinguishable*	pt004827m	0	6	0	6
	pt005089m	*Indistinguishable*	pt004827m	0	2	0	2
	pt005064m	*Indistinguishable*	pt004827m	0	2	0	2
	pt004830m	*Distinct*	pt004830m	12	1	1	14
	pt039322m	*Differentiable*	pt004830m	1	1	0	2
	pt004839m	*Distinct*	pt004839m	11	0	1	12
	pt012284m	*Distinct*	pt012284m	2	0	1	3
	pt020991m	*Differentiable*	pt020991m	2	0	3	5
	pt030211m	*Indistinguishable*	pt021000m	0	5	0	5
	pt021000m	*Indistinguishable*	pt021000m	0	5	0	5
	pt021032m	*Distinct*	pt021032m	1	0	0	1
	pt023595m	*Distinct*	pt023595m	6	0	3	9
	pt023595m	*Distinct*	pt023595m	6	0	3	9
	pt033373m	*Differentiable*	pt033373m	2	1	2	5
	pt033727m	*Indistinguishable*	pt033373m	0	1	2	3
CAld5H	pt032677m	*Differentiable*	pt032677m	3	0	6	9
	pt025189m	*Differentiable*	pt025189m	7	0	5	12

Table 3.2 continued:

Protein Family	Protein	Protein Parsimony	Seed	DU peptides	PGU peptides	NU peptides	Total peptides
COMT	pt000701m	*Indistinguishable*	pt000702m	0	4	15	19
	pt000702m	*Indistinguishable*	pt000702m	0	4	15	19
	pt010103m	*Distinct*	pt010103m	3	0	0	3
	pt015982m	*Distinct*	pt015982m	28	0	15	43
	pt018020m	*Indistinguishable*	pt018020m	0	5	0	5
	pt018431m	*Indistinguishable*	pt018020m	0	5	0	5
	pt020855m	*Indistinguishable*	pt020853m	0	3	0	3
	pt022214m	*Indistinguishable*	pt020853m	0	3	0	3
	pt020853m	*Indistinguishable*	pt020853m	0	3	0	3
	pt020964m	*Indistinguishable*	pt020853m	0	2	0	2
CAD	pt003155m	*Distinct*	pt003155m	1	0	0	1
	pt003292m	*Distinct*	pt003292m	11	0	0	11
	pt004002m	*Distinct*	pt004002m	2	0	0	2
	pt004753m	*Distinct*	pt004753m	31	0	2	33
	pt018077m	*Indistinguishable*	pt018073m	0	3	0	3
	pt018073m	*Indistinguishable*	pt018073m	0	3	0	3
	pt039056m	*Distinct*	pt039056m	8	0	0	8

*A detailed classification of the peptides detected within 10 protein families contributing to lignin biosynthesis. A protein was marked as distinct, differentiable, or indistinguishable according to its number of database-unique (DU) peptides detected. After reorganizing proteins into their protein groups, peptide uniqueness was reevaluated for protein-group uniqueness (PGU). The number of non-unique (NU) peptides was also reported. The following protein families were observed: phenylalanine ammonia-lyase (PAL), cinnamate 4-hydroxylase (C4H), 4-coumarate CoA ligase (4CL), p-hydroxycinnamoyl-CoA: quinate shikimate p-hydroxycinnamoyltransferase (HCT), p-coumarate 3-hydroxylase (C3H), caffeoyl-CoA O-methyltransferase (CCOMT), hydroxycinnamyl-CoA reductase (CCR), coniferaldehyde 5-hydroxylase (CAld5H) caffeate O-methyltransferase (COMT), and cinnamyl alcohol dehydrogenase (CAD).

Thus, by not anticipating the presence of these peptides, database search algorithms are more likely to interpret fewer peptide-derived MS/MS spectra when analyzing proteomes of higher eukaryotes. Reanalysis of unassigned tandem mass spectra was performed to determine the magnitude of peptide-derived spectra that remained unmatched to a sequence, thereby providing the proportion of "missing" peptide identifications in a run.

To compare the rates of peptide-spectrum matching (PSM) between eukaryotes and prokaryotes, we contrasted MS/MS data from *Populus* with a simpler bacterium, *Escherichia coli*[209]. In both cases, proteolytic peptides were measured on the same instrument using identical methods to minimize experimental biases. The instrumental acquisition and chromatographic distribution of all MS/MS spectra collected were similar for both organisms (Figure 3.5A). However, the ability to successfully match experimental MS/MS spectra to theoretical database sequences was superior in *E. coli*. A greater percentage (86%) of *Populus* MS/MS spectra remained unassigned, as compared to only 63% of the MS/MS spectra collected for *E. coli*. A closer look at the proportion of unassigned peptide-derived spectra was used to determine if the observed discrepancies in peptide identifications could be attributed to the incompleteness of the reference database. Spectral quality assessment was used to identify the number of unassigned high-quality spectra, i.e., a population of spectra that likely represents mutated, modified or novel peptides. A conservative set of criteria, based on previous implementations of spectral analysis was utilized in the assessment of MS/MS spectral quality[210-211]. A spectrum was considered high quality if the parent charge state was calculated to be greater than +1 and if the spectrum contained three or more peaks within 20% of the base peak intensity with a minimum intensity of 2,500 counts. Using this approach, we performed an assessment of MS/MS spectra quality to distinguish high-quality unassigned spectra from low-quality unassigned spectra in the representative MS runs from *Populus* and *E. coli*. Spectra analysis revealed that, of the total MS/MS spectra collected for *Populus* and *E. coli*, the percentage of high-quality MS/MS spectra (45%) within the representative MS run for *Populus* contained almost twice the percentage (24%) in the *E. coli* run (Figure 3.5B).

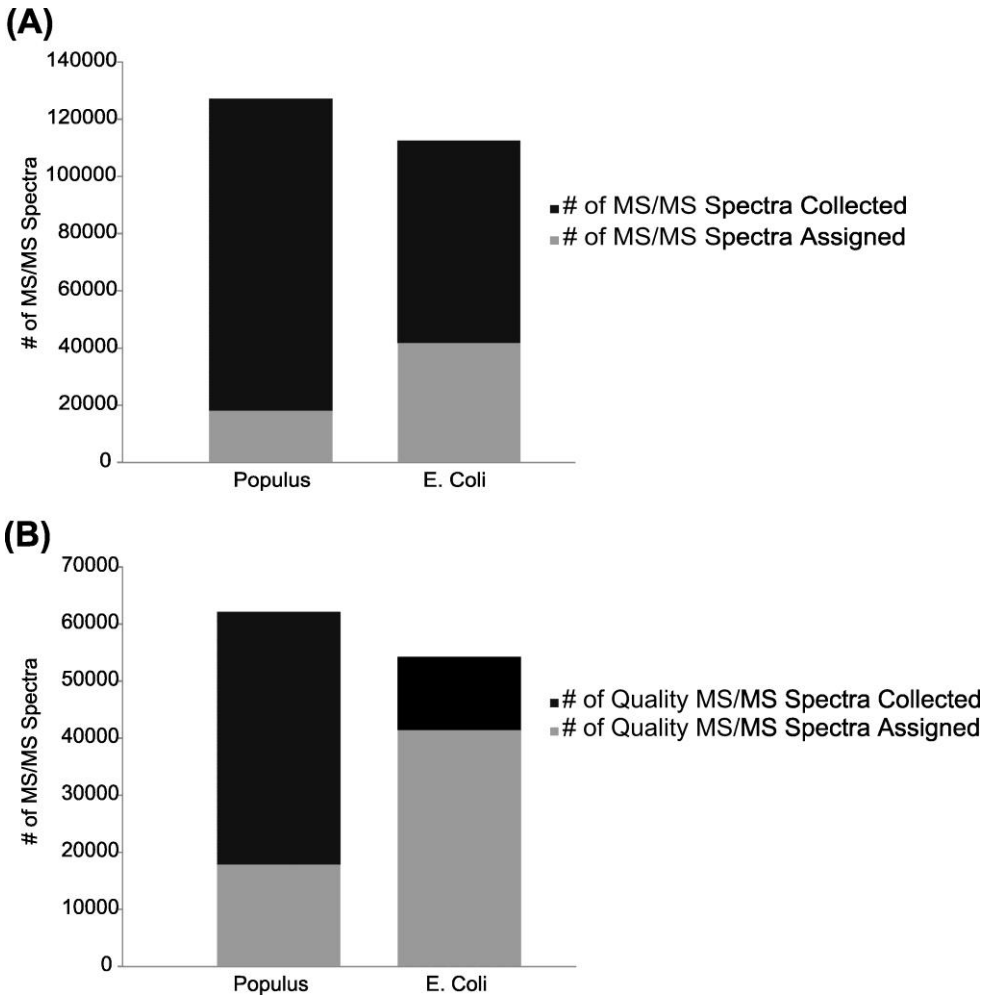


Figure 3.5. Spectral quality assessment distributions for *Populus* and *E. coli*. A) Comparison of peptide-spectrum matching rates between *E. coli* and *Populus*. B) Quantitative assessment of the proportion of high-quality MS/MS spectra collected versus those assigned for the representative MS runs from *E. coli* and *Populus*.

Nonetheless, the ability to successfully match the high-quality experimental MS/MS spectra to database sequences remained more common in *E. coli*. A greater percentage of *Populus* high-quality MS/MS spectra (77%) remained unassigned, as compared to only 45% of the high-quality MS/MS spectra collected for *E. coli*. This suggests a critical need to evaluate bioinformatic approaches to rescue the lost, high-quality spectra.

3.3.2 Single Amino Acid-Resolved Populus Proteomics

One source of unassigned high-quality tandem MS spectra may be peptides containing SAAPs. *Populus* has an estimated one SNP per 200 base pairs[114], while humans have an estimated one SNP per 1.9 kilobase pairs[212]. The biological implication of a SNP depends on its positional location within the genome and gene structure. Within coding regions, a SNP can be either synonymous which does not alter the amino acid or nonsynonymous which results in an amino acid substitution. Detection of SAAPs not only identifies amino acid changes that have physiochemical consequences but also reveals information regarding sequence, and perhaps phenotypic, variability within a proteome. Therefore, a database containing SNP-based SAAPs and other sequence variations could be highly informative.

To explore the prevalence of SAAPs, a single MS run from within the 60 described above was searched against an expanded *Populus* database that included a list of tryptic peptides generated from predicted SAAP variants in the database. In brief, high-throughput SNP discovery through deep (30X depth per genotype) resequencing of 19 trees yielded 16 million SNPs in the *Populus* genome (485 Mb) (unpublished results). For this analysis, a subset of these SNPs present in 2 *P. trichocarpa* and 2 *P. deltoides* genotypes were considered. Of the 17 million amino acid positions found in *P. trichocarpa's* 45,778 protein-coding gene models, ~400,000 amino acid positions due to non-synonymous SNPs (SAAP) were investigated. All possible combinations of SNP-influenced peptides (SAAP peptides) were predicted and subjected to in silico tryptic cleavage using PeptideSieve software with the following parameters: maximum mass criterion of 5000, minimum sequence length of 6, maximum sequence length of 50 and allowing for 4 missed cleavages. Some of the non-synonymous amino acid changes resulted in new tryptic cleavage sites or resulted in disappearance of these sites. These

were taken into consideration while predicting the peptides. To detect the expression of a SAAP peptide, experimental MS/MS spectra from one MS run were compared to theoretical tryptic peptide sequences generated from a target database consisting of the protein database *of P. trichocarpa* (v2.0) and all predicted SAAP peptides. Each SAAP peptide was concatenated to the target database as a new protein entry, in which ten tryptophan residues flanked both sides of the peptide sequences. For SAAP peptides that originated from the N-terminus of a protein, the tryptophan residues were excluded from the beginning of the SAAP peptide. Similarly, for each SAAP peptide that originated from the C-terminus of a protein, the tryptophan residues were excluded at the end of the SAAP peptide%. With the high frequency of SAAPs in *Populus*, over 700,000 distinct SAAP positions and 7,200,000 new peptides were included in our database. All MS/MS were searched with SEQUEST and filtered by DTASelect as described previously. Once peptide-spectrum matches were identified, filtering criteria were controlled to yield peptide FDRs less than 1%. We found that *Populus* proteins on average contained 17 SAAPs. When identifying SAAPs from MS/MS spectra, it is important to differentiate these from post-translational modifications (PTMs) or peptide modifications generated during sample processing that result in mass shifts which are isobaric to several amino acid substitutions. For example, the covalent addition of a methyl group to a K, R, E, or Q produces a mass shift that is similar to the following amino acid changes: D to E, S to T, V to I/L, and G to A. Therefore, all spectra interpreted as both a PTM and a SAAP were discarded to lower the identification of false positives. To identify a targeted common set of PTMs, MS/MS spectra were analyzed by an automated software tool, InSpecT[55], at a peptide FDR of 2%. In total, 271 spectra that matched to both a PTM and a SAAP peptide were removed from the analysis. Using conservative search criteria, we were able to identify a total of 1,354 peptides containing a SAAP and 201 peptides that become tryptic due to a K or R substitution. Although the new SAAP peptides account for 2% of high-quality unassigned spectra, these newly identified peptides correspond to 502 proteins. Among these, we identified 97 proteins that had not been previously identified. Interestingly, for those proteins containing a SAAP peptide, their overall peptide coverage increased by an average 25%.

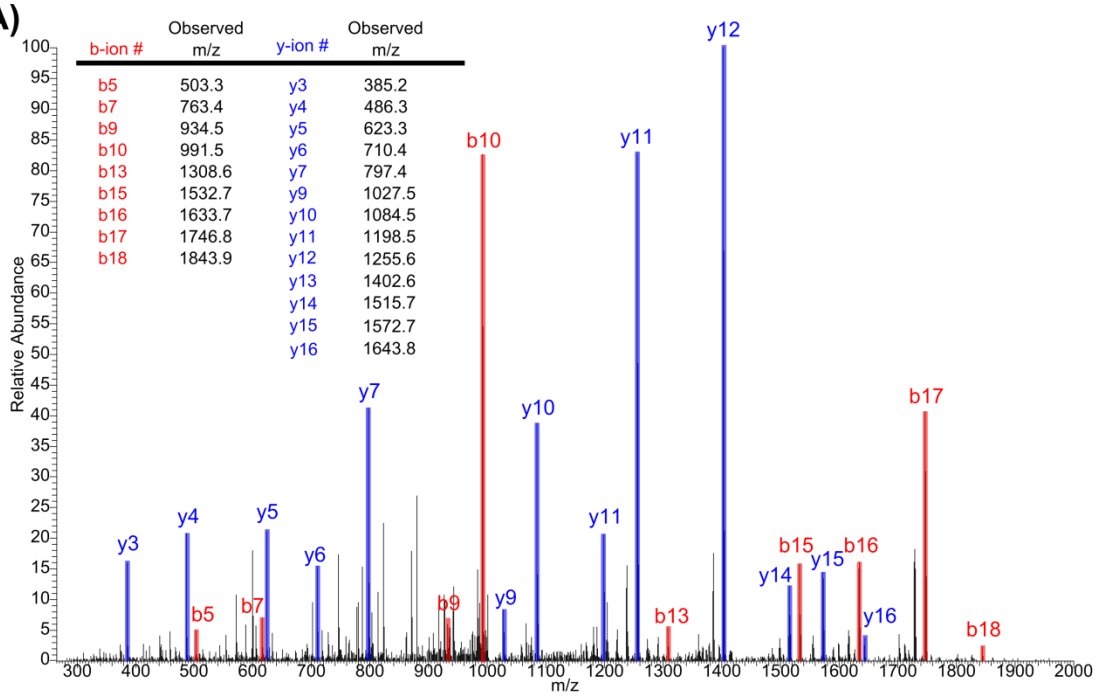
Due to the widespread distribution of SAAP peptides in the database, it seems probable that the detected SAAP peptides would map randomly across the proteome. However, our data suggests that the detected population of proteins containing a SAAP peptide map to specific and functionally similar groups. Grouping the SAAP proteins into KOGs, the vast majority of SAAP proteins belonged to the four specific functional categories: unknown function, signal transduction, post-translational modification, and carbohydrate transport and metabolism. Although these functional categories are among the most abundant categories in phloem and xylem, we note that other abundant functional categories, such as general function and translation, do not contain a large number of proteins containing SAAPs. Therefore, it appears that the overrepresentation of non-synonymous substitutions for the aforementioned functional categories is not a result of their expression levels, but rather that these proteins are under low selective pressure. Although it is unclear how many of these proteins represent evolutionary novelties, future comparative proteomics studies may identify expression patterns that reveal the outcomes of such mutations. In some instances, the location of these mutations could compromise or benefit an enzyme: replacing catalytic, binding, or substrate determining residues with amino acids differing in size, polarity, or hydrophobicity can either disrupt or modulate the activity of an enzyme.

For example, when looking at the monolignol biosynthesis pathway, we identified a SAAP within phenylalanine ammonia lyase (PAL), the entry enzyme into the phenylproponoid pathway. As shown in Figure 3.6, a mass shift of +1 Da and the experimental b- and y- ion fragmentation pattern coincides with the predicted SAAP substitution of an asparagine (N) with an aspartic acid (D) at position 138.

Figure 3.6. Single amino acid polymorphism-resolved peptide identification in PAL. A) MS/MS spectra of the genomic tryptic peptide (FLNAGIFGNGTESSHTLPR) and the B) SAAP tryptic peptide (FLNAGIFGDGTESSHTLPR). C) A partial amino acid (single letter codes) sequence alignment *of P. trichocarpa* (PtPAL) with other members of the phenylalanine ammonia-lyase family (PcPAL, *P. Crisum* and AtPAL, *A. thaliana*). Only the region near the SAAP-containing peptide is shown. The yellow box highlights the substrate specificity residues and the green box highlights the SAAP position.

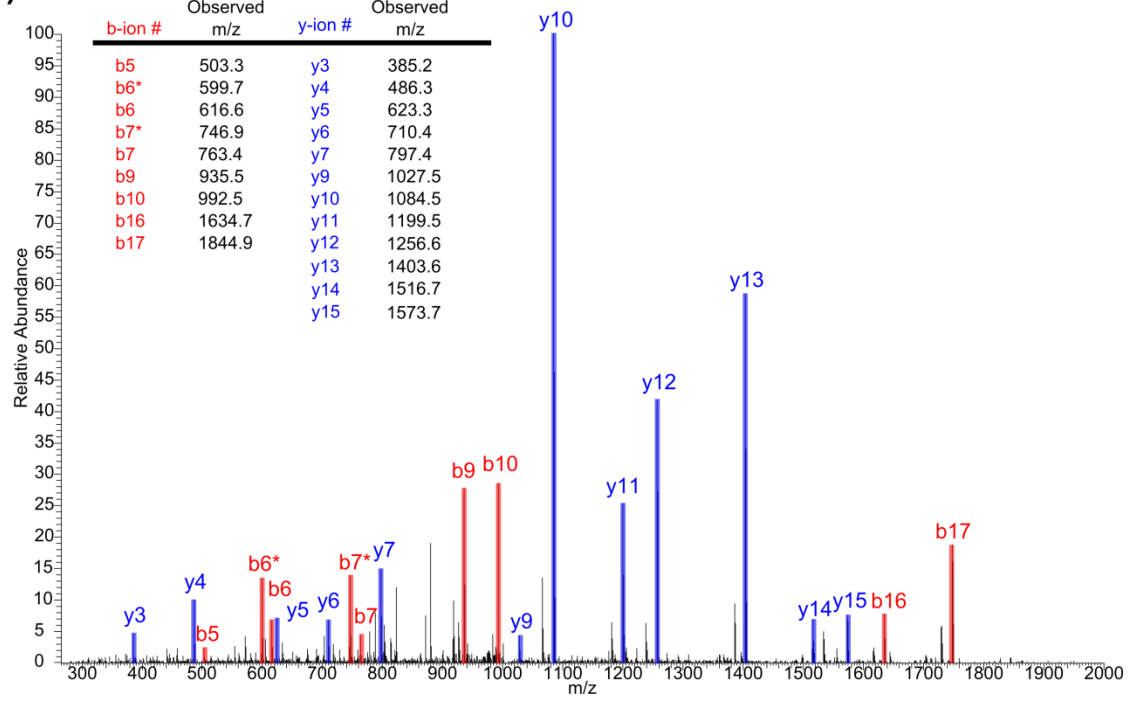
(A)

b-ion #	Observed m/z	y-ion #	Observed m/z
b5	503.3	y3	385.2
b7	763.4	y4	486.3
b9	934.5	y5	623.3
b10	991.5	y6	710.4
b13	1308.6	y7	797.4
b15	1532.7	y9	1027.5
b16	1633.7	y10	1084.5
b17	1746.8	y11	1198.5
b18	1843.9	y12	1255.6
		y13	1402.6
		y14	1515.7
		y15	1572.7
		y16	1643.8



(B)

b-ion #	Observed m/z	y-ion #	Observed m/z
b5	503.3	y3	385.2
b6*	599.7	y4	486.3
b6	616.6	y5	623.3
b7*	746.9	y6	710.4
b7	763.4	y7	797.4
b9	935.5	y9	1027.5
b10	992.5	y10	1084.5
b16	1634.7	y11	1199.5
b17	1844.9	y12	1256.6
		y13	1403.6
		y14	1516.7
		y15	1573.7



(C)

```
PtPAL          DSYGVTTGFGATSHRRTKQGGELQKELIRFLNAGIFGNGTESSHTLPRSATRAAMLVRIN
PtPAL (SAAP)   DSYGVTTGFGATSHRRTKQGGELQKELIRFLNAGIFGDGTESSHTLPRSATRAAMLVRIN
PcPAL          DSYGVTTGFGATSHRRTKQGGALQKELIRFLNAGIFGNGSD--NTLPHSATRAAMLVRIN
AtPAL          DSYGVTTGFGATSHRRTKNGV ALQKELIRFLNAGIFGSTKETSHTLPHSATRAAMLVRIN
```

While the effect of the observed polymorphism is unknown, the localization of the substitution within a few amino acids of the substrate-binding site may impact the binding of coumarate to the substrate specificity residues[213]. Because studies have shown that PAL serves as a regulatory control point for the entire pathway[214], any mutations compromising or altering the activity of the enzyme will, in fact, impact the overall lignin content.

3.4 Conclusions

While it is still unknown what percent of the *Populus* proteome is expressed given a specific time and tissue, combining tandem mass spectra from 60 MS runs yielded a preview of protein expression in xylem and phloem. Perhaps one of the most challenging tasks in proteomic studies of higher eukaryotes is inferring which proteins are present in a particular sample based on the observed peptides. An enhanced bioinformatic workflow alleviated some of the difficulties associated with data interpretation by recasting protein identifications as protein groups, which have a high degree of sequence similarity and therefore most likely share similar biological roles.

In addition, to fully characterize the boundaries of assignable peptides, we assessed spectral quality and found a large portion of the high-quality spectra remained unassigned. When dealing with higher eukaryotes such as plants, a major issue for tandem MS and peptide identification algorithms is the high level of sequence variation, including naturally occurring PTMs and SNP-based SAAPs. The exact scope and frequency of these detectable protein variants has, to our knowledge, never been reported to date in any plant. By investigating the prevalence of detectable SAAPs, we provide a glimpse of detectable proteins beyond the 'basic' proteome (predicted gene products). All together, the integration of deep proteome measurement on an extensive sample set with protein clustering and identification of protein sequence variants pioneered a level of proteome characterization for *Populus* that has not been possible before.

Although the resulting data set provided a more accurate and informative perspective that allowed us to characterize the landscape of protein expression in xylem and phloem, the experimental workflow can be further improved by addressing some key

analytical challenges posed by plant cells. As discussed in the next chapter, our pre-existing sample preparation (i.e., non-detergent-based lysis and protein extraction), which was developed for microbial samples, and mass spectrometer instrumentation did not provide optimal analytical depth and comprehensiveness required to accurately answer biological questions pertinent to *Populus* research.

CHAPTER 4

DEVELOPING AN EXPERIMENTAL STRATEGY FOR *POPULUS*: THE INTEGRATION OF A DETERGENT-BASED LYSIS PROTOCOL AND THE DUAL-PRESSURE LINEAR ION TRAP MASS SPECTROMETER

All of the data presented below has been adapted from the following published journal article:

Paul Abraham, Richard Giannone, Rachel Adams, Udaya Kalluri, Gerald Tuskan, Robert Hettich (2012). "Putting the Pieces Together: High-performance LC-MS/MS Provides Network-, Pathway-, and Protein-level Perspectives in Populus". *Molecular and Cellular Proteomics 2013* 12(1): 106-119. Sample preparation and mass spectrometry experiments were performed by Paul Abraham. Biological data analysis was performed by Paul Abraham and Richard Giannone.

4.1 Introduction to Experimental Challenges

Mass spectrometry (MS)-based proteomics has experienced tremendous growth in recent years, leading to the establishment of numerous protocols, platforms, and workflows for the characterization of protein expression at the genome level[215]. While these advancements have facilitated comprehensive proteomic investigations of simple bacterial isolates and microbial communities, the application of MS-based proteomics for plants and other higher eukaryotes remains underdeveloped. Recently, large-scale proteomic studies have been directed at characterization of *Populus*, a woody perennial model organism. With the recent release and subsequent curation of the *P. trichocarpa* genome, these large-scale MS-based proteomic investigations offer the potential to introduce new biological insights into woody perennial plant biology[216]. As shown in Chapter 3, we demonstrated the ability to measure ~17% of the *Populus* proteome by coupling multi-dimensional liquid chromatography (MudPIT) with nano-electrospray tandem mass spectrometry (2D-LC-MS/MS)[217]. Relative to the two-dimensional gel-based approaches[218], MudPIT provides enhanced separation and when used in conjunction with MS/MS, surpasses the throughput and number of identifiable proteins detected in complex mixtures[84]. Although we have demonstrated the general

effectiveness of this approach, the identification and quantitation of the proteins expressed in a plant cell or tissue are still notoriously complicated by a number of factors, including the size and complexity of plant genomes, abundance of protein variants, as well as the dynamic range of protein identification. To overcome these challenges, improvements are needed in sample preparation and MS instrumentation.

The architecture of plant cell walls provides resistance to chemical and biological degradation, thus requiring mechanical and detergent-based lysis for optimal proteome analysis. However, this criterion presents a major challenge for plant proteomic research using electrospray mass spectrometry, as detergent-containing solutions can impede enzymatic digestion and cause significant analyte suppression[219]. Therefore, most plant proteomic studies using the 'MudPIT' strategy apply mechanical disruption in conjunction with a detergent-free preparation method[220]. Typically, strong chaotropic agents such as urea and guanidine hydrochloride are used for the extraction, denaturation, and digestion of proteins. In a recent study, Mann and colleagues introduced a filter-aided sample preparation (FASP) method that utilizes and effectively removes sodium dodecyl sulfate (SDS) prior to enzymatic digestion and electrospray analysis[221]. This study demonstrated enhanced retrieval of peptides from biological materials, yielding a more accurate representation of the proteome. We developed a similar experimental approach for extraction of proteins from plant tissue in order to obtain a more comprehensive, unbiased proteome characterization well beyond that achievable with currently available methods. Similar to the FASP method, we demonstrate the power of SDS for proteomic sample preparation, not only in its ability to more-thoroughly lyse cells, but also its ability to better solubilize both hydrophilic and hydrophobic proteins. This powerful attribute gives proteolytic enzymes maximum opportunity to generate peptides specific to their cleavage potential so that at least a few representative peptides can be obtained for proteins that would have otherwise been discarded or lost due to insolubility, e.g., membrane-bound proteins. Rather than performing a buffer exchange with urea, depletion of SDS is achieved by precipitating proteins out of solution using trichloroacetic acid.

Characterization of protein expression in plants is further complicated by the heterogeneous mixture of various cell types, each with a unique proteome signature and individualized response to environmental chemical or physical signals. This inherent complexity of plant proteomes and the large dynamic range in protein abundance overwhelms current analytical platforms[222]. Moreover, biochemical regulatory networks in plants are more elaborate and dynamic than in microbial species; consequently, many biological components are left undiscovered, including modified peptides and low-abundance proteins[79, 223-224]. Recent developments in ion-trap MS instrumentation, namely the dual-pressure linear ion trap mass spectrometer (LTQ Velos), have demonstrated improved ability to comprehensively characterize complex proteomics samples[225]. Featuring a newly designed ion source and a two-chamber ion trap mass analyzer, the LTQ Velos achieves greater dynamic range, sensitivity, and speed of spectral acquisition when applied to complex proteomic samples. Cumulatively, the technological advancements afford substantial increases in the detection and identification of both proteins and unique peptides when compared to existing state-of-the-art technologies. Therefore, to satisfy the need for depth of proteome characterization in plants, we coupled the detergent-based method with the newly developed LTQ Velos for mass spectrometry measurements of the *Populus* proteome.

4.2 Global Protein Identification in *Populus*

A protein sample derived from plant tissue is likely to consist of over 10,000 different protein species present at any time and thus the complexity far exceeds an analogous sample derived from any prokaryotic species. The first step in accurate and deep proteome characterization in these mixtures must consist of an optimal cell lysis and protein solubilization strategy. For plant tissue, we devised a bottom-up proteomic workflow that combines the advantages of extensive proteome solubilization in SDS with the benefits of in-solution digestion.

In an effort to generate a high-density proteomic atlas that accurately captures the predicted *Populus* proteome, individual proteome maps of the four major organ-types were integrated. In total, we performed multiple (5-6 each) LTQ Velos ion-trap mass

spectrometry measurements on proteome extracts from root, stem and both mature (fully expanded, leaf plastichronic index (LPI) 10-12) and young leaf (LPI 4-6) samples. The resulting tandem mass spectra (MS/MS) were searched (SEQUEST) against the most recent protein database of *P. trichocarpa*, containing 45,778 predicted proteins and supplemented with the chloroplast and mitochondrial proteomes.

In plants, the task of assigning identified peptides to their respective proteins is not trivial. Due to the peptide-centric nature of shotgun proteomics, peptides that map to multiple proteins in a reference database can lead to ambiguous identifications. Within higher eukaryotes, this imposes a considerable challenge because shared or degenerate peptides, which result from segmental duplications, homologous proteins or splicing variants and comprise a large fraction of total extracted peptide library[184-185]. To date, there are different methods for aggregating MS evidence for protein assembly[73]. As discussed in Chapter 3, the most advantageous framework to classify and validate protein identifications in higher eukaryotes should include the following: 1) a means to report the minimum of proteins implicated by at least one unique peptide and 2) the ability to account for database redundancies by clustering similar proteins into groups by sequence homology.

Using the principle of parsimony with Occam's razor constraints, 7,720 *Populus* proteins were confidently identified (classified as distinct or differentiable), and 4,520 proteins were categorized as indistinguishable. Although widely used, the guidelines in the suggested nomenclature make data interpretation more complicated and less accurate, especially in highly redundant proteome databases like *Populus*.

For this reason, we proposed a strategy that incorporates additional supporting information (i.e., sequence homology) to better infer the existence of proteins. While this approach can be applied to bottom-up proteomic studies of plants in general, it confers demonstrable advantages for *Populus* specifically. Proteins sharing 90% or more sequence identity within the *Populus* database were clustered into groups. Each protein group was defined by a single representative protein sequence called a seed, where each seed shares ≥90% sequence identity with all other members of that group. Observed peptides from the originally searched protein entries were then directly referenced back to

the clustered database. For the current data set that included 63,056 tryptic peptides, ~25% were previously shared within the original *Populus* database (non-unique/degenerate) but were reclassified as unique to a particular protein group in the newly constructed database. This illustrates the advantage of implementing a "protein group-centric" approach, such that including information about sequence homology allows the interpreter to readily assess the relatedness between shared peptides of indistinguishable proteins derived from gene duplication and splice variants. Moreover, as clustered proteins are ≥90% similar to one another, members of a particular group likely exhibit similar functional roles which, when applied to semi-quantitative proteomics, allows for a more robust analysis of functional signatures across conditions, time points or organ types. In other words, this strategy effectively reduces the complexity of the functional analysis and biological interpretation of plant data.

Based on this approach, a total of 11,692 protein assignments across all organ-types were reduced into 7,538 protein groups at an average false-discovery rates of <1% at the peptide level. Protein groups were populated by as many as 21 members, with one-membered groups (i.e., singletons) representing only 36% of the total. In total, we were able to measure 25% of the predicted proteins for *Populus*. Generating complete proteome maps of higher organisms is a difficult task as it is unlikely the entire ensemble of polypeptide species encoded by a genome will be expressed at any given time. Nevertheless, this integrated data set provides an "information backbone" that captures baseline protein expression across spatially and functionally distinct pathways. This holistic view of plant-wide protein expression will provide a better understanding of the detected components (i.e., proteins, pathways, etc.) in the context of relationships between organs.

4.3 Depth of Analysis of the *Populus* Proteome

Having established robust peptide/protein identification criteria, we sought to assess the depth of our data set by four critical figures of merit, proteome sequence coverage, sensitivity, data acquisition speed, and dynamic range. Despite differences in organ background, similar total protein sequence coverage (median=19%) was achieved

(Figure 4.1), a value comparable to recent work employing a similar approach to analyze yeast[226]. Of the four organ-types, the mature leaf proteome consisted of proteins with lower total sequence coverage. Concomitantly, there were fewer proteins with high sequence coverage. We speculate that the heterogeneity of the expected protein population expressed in mature leaf (i.e., membrane-related proteins, post-translational modifications, etc.) is perhaps less suited for the current trypsin-based schema. Transmembrane prediction using Hidden Markov Models (TMHMM)[227] analysis revealed similar identification rates of proteins with transmembrane domains (6-7% across all organs), suggesting that the systematic decrease in protein sequence coverage is more likely due to changes in the frequency of post-translational modifications or some other phenomena related to the types of proteins being expressed.

Unlike the experimental design presented in Chapter 3, we incorporated a detergent-based lysis and protein solubilization strategy to improve the overall sensitivity of our proteome analysis. Plant cells contain a relatively high level of proteins whose extraction is rendered difficult because they are associated with the cell wall and other compounds, such as storage polysaccharides, phenolic compounds, and lipids. In general, these proteins are poorly soluble in aqueous and chaotropic solvents, and therefore efficient protein extraction of all detectable proteins requires a detergent-based approach. Although SDS is routinely used as the reagent of choice, this ionic detergent can, even in small concentrations (~0.1%), preclude enzymatic digestion of proteins and can cause significant ion suppression in electrospray. Therefore, TCA precipitation was integrated to reduce the concentration of SDS to well below threshold levels, allowing better proteome coverage without interferences. To compare the two methodologies, a population of plant cells was processed with each method and analyzed by the same mass spectrometer. As shown in Figure 4.2, the SDS-based approach proved superior to the previous methodology with regards to the detection of low-abundant proteins. In addition, the SDS-based method demonstrated remarkable gains in protein identifications (+57%), peptide identifications (+48%), and the number of spectra counts (+78%).

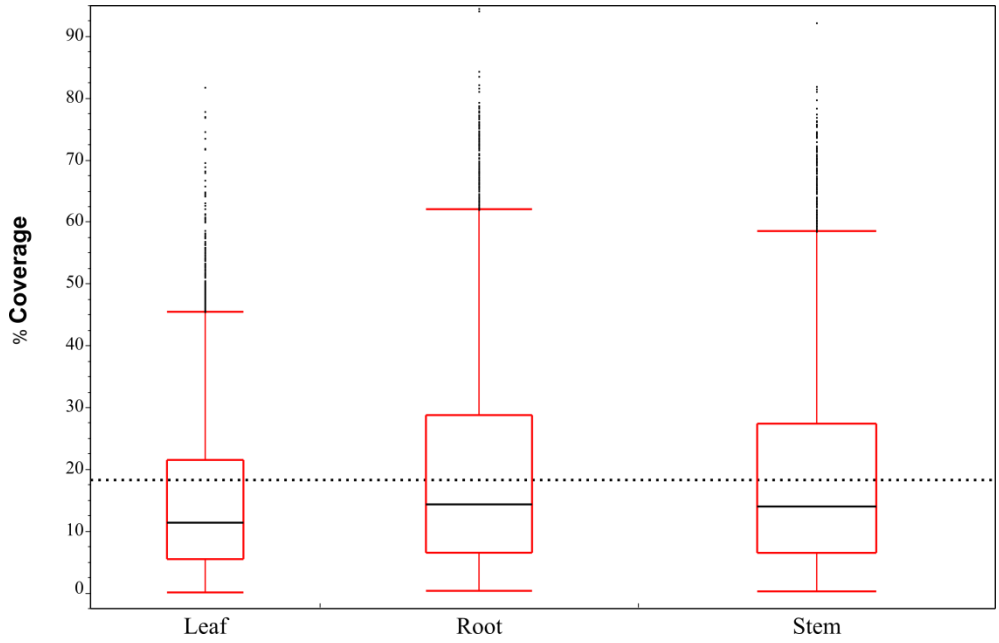


Figure 4.1. Box and whisker plot of total percent sequence coverage in Populus leaf, root, and stem proteomes. The distribution of the total sequence coverage per protein by organ is displayed as a box and whisker plot. Data falling within the interquartile range (25% - 75%) is boxed with the median value highlighted by the horizontal line, with the overall mean represented by a dashed-horizontal line. Predicted outliers fall above the upper horizontal line. Sequence coverage data distributions are synonymous for root, stem, and young leaf. Sequence coverage for mature leaf is systematically depressed by roughly 5%, perhaps indicating enrichment of proteins with transmembrane domains or those with a higher-degree of post-translational modification.

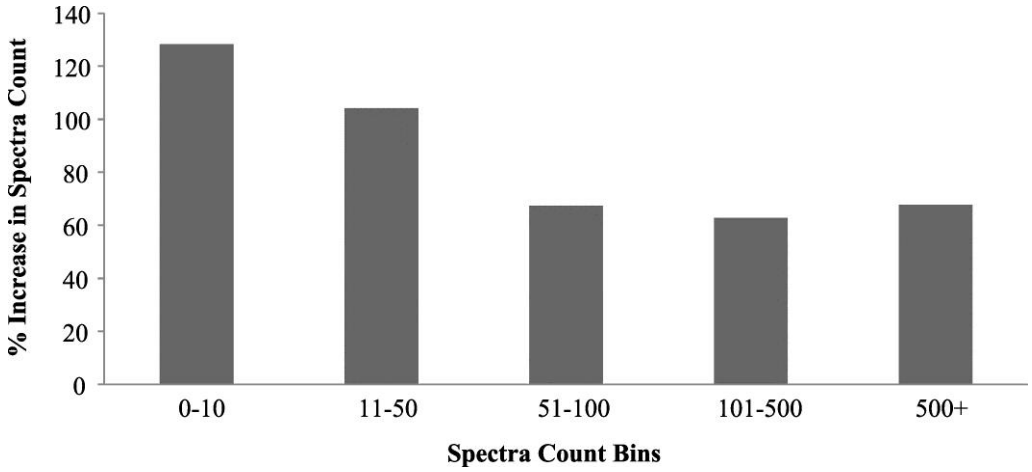


Figure 4.2. A comparison between the SDS-based and non-detergent-based methodologies. Each spectra count bin represents a degree of protein abundance, where the "0-10" bin represents least abundant proteins and the "500+" bin represents the most abundant proteins observed. The size of each bin is dictated by the % gain in spectra count when using the SDS-based approach.

Electrospray ionization presents the mass spectrometer with a dynamic population of peptides, of which only a fraction is selected for sequencing[228]. Consequently, highly abundant peptides limit the sampling and identification of low-abundant peptides. Because the LTQ Velos platform includes advances that benefit the analysis of low-abundant and low signal-to-noise precursors, we sought to quantitatively assess the capabilities of the instrument. By comparing our current data set against our previous in-depth Chapter 3, which used the LTQ XL platform, we examined the achievable depth of proteome characterization. When examining the distribution of the identified precursor ions versus local signal-to-noise ratios, the LTQ Velos platform increased the identification of low signal-to-noise precursor ions compared to the LTQ XL (Figure 4.3). Furthermore, peptide populations created from complex mixtures often tax the sequencing speed of MS instruments such that the mass spectrometer is incapable of targeting every eluting peptide and thus misses "sequenceable" peptides. As anticipated, the faster acquisition speed facilitated a 2-fold increase in the number of scans collected and assigned as well as the total number of proteins identified (Figure 4.4A-B). Given these improvements provided by the Velos platform, we anticipated a sizeable increase in the analytical dynamic range. Indeed, protein dynamic range spanned 5-6 orders of magnitude, representing a 1-2 order of magnitude increase when compared to the LTQ XL platform (Figure 4.4B). Together, these increases in sensitivity and speed augment the analytical dynamic range, providing demonstrably better depth of proteome characterization.

Similar experimental strategies directed towards deep proteome coverage in eukaryotes, like yeast, have measured a remarkably large dynamic range of protein expression[229]. Unlike yeast[98], there is no available information regarding known cellular concentrations (copies/cell) for proteins spanning the entire abundance range in *Populus*. Therefore, it is a challenge to accurately assess the biological dynamic range achieved by this approach. Nevertheless, the experimental and dual-pressure ion trap designs include substantial improvements that benefit the analysis of complex protein mixtures.

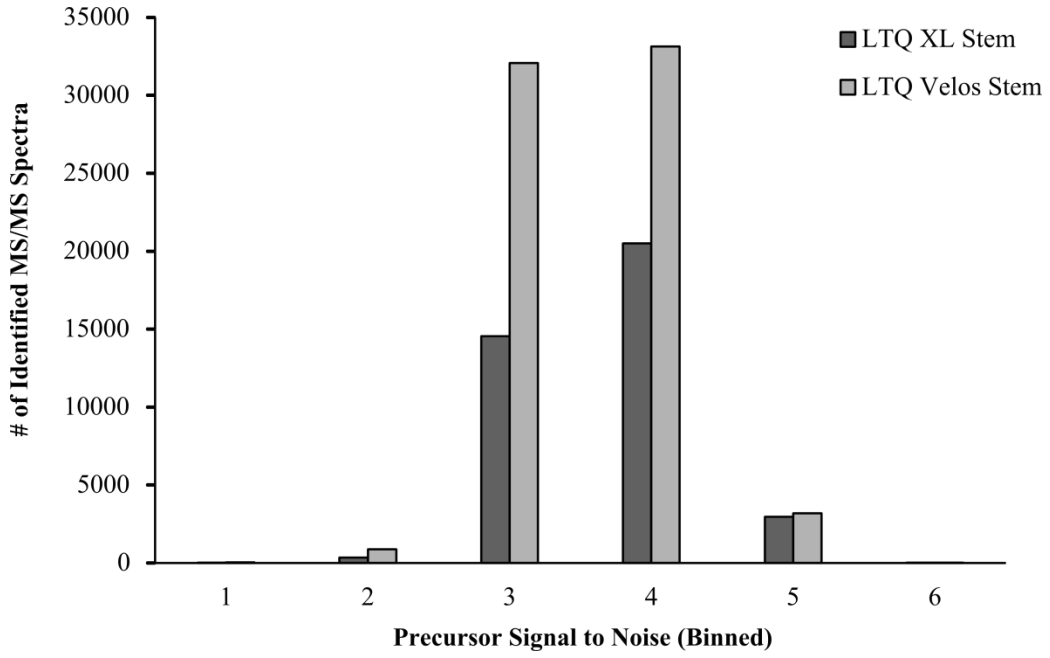
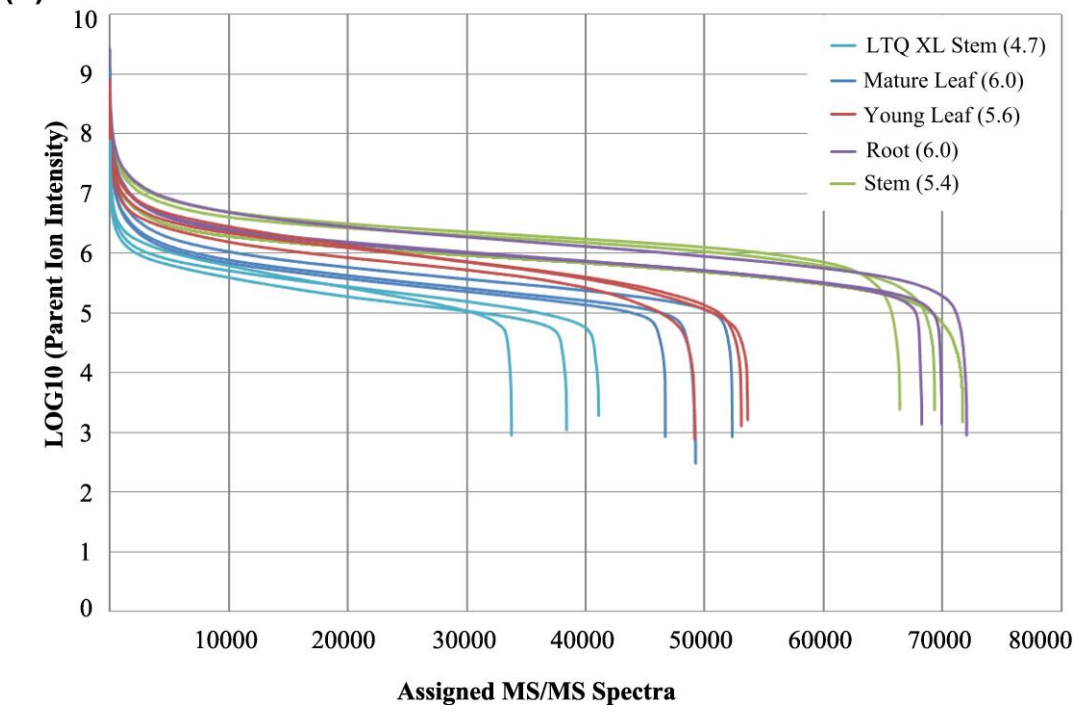


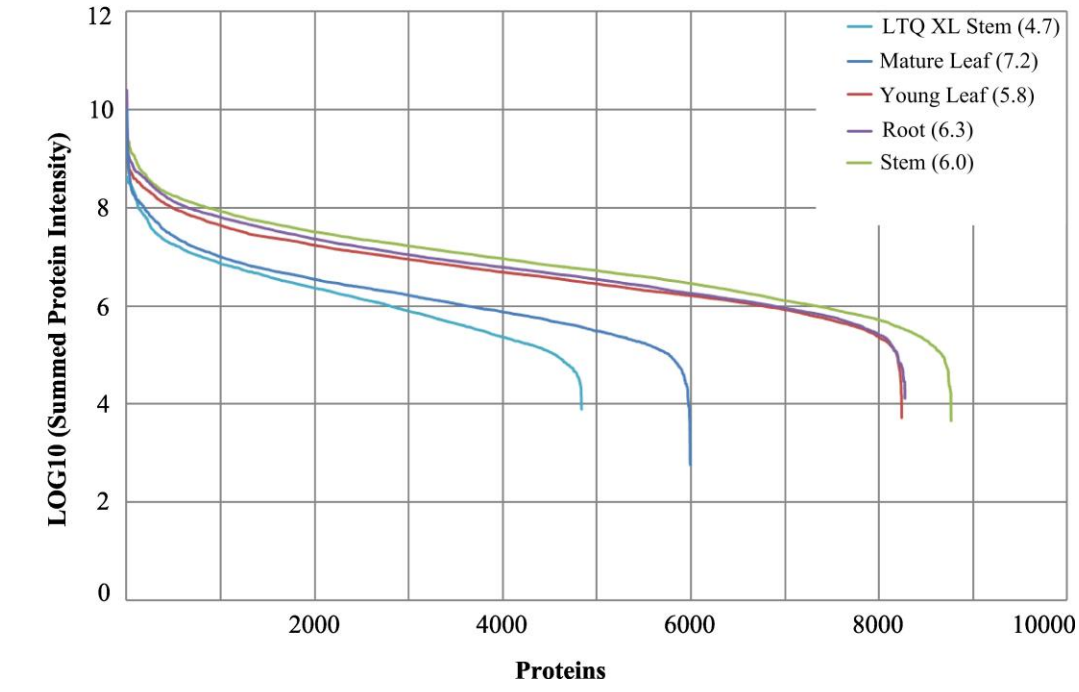
Figure 4.3. A comparison of signal-to-noise ratio of identified peptides between the LTQ XL and LTQ Velos platform. Number of identified MS/MS events versus the estimated local signal-to-noise ratios of precursor ions. Overall, the LTQ Velos spent most of the extra 2X sampling time identifying more low abundance precursors than LTQ XL.

Figure 4.4. Peptide and protein dynamic range in *Populus*. Dynamic range of measurement was assessed for each identified peptide and protein across three replicate runs for all four organ types. (A) Maximum ion intensity values obtained for each peptide's extracted ion chromatogram (y-axis) were ranked by intensity and plotted against cumulative number of assigned MS/MS spectra (x-axis). Curves represent individual replicates per organ to identify run-to-run differences in dynamic range. (B) Assembled protein intensity (y-axis), calculated by summing constituent peptide intensities across all replicates, was plotted against the cumulative number of identified proteins (x-axis) to identify the overall protein dynamic range achieved per individual organ. Dynamic range values, represented as magnitudes (base 10), are listed in the figure legend. Light blue: LTQ stem; blue: LTQ-Velos mature leaf; red: LTQ-Velos young leaf; purple: LTQ-Velos root; green: LTQ-Velos stem.

(A)


LOG10 (Parent Ion Intensity)
Assigned MS/MS Spectra
LTQ XL Stem (4.7)
Mature Leaf (6.0)
Young Leaf (5.6)
Root (6.0)
Stem (5.4)

(B)


LOG10 (Summed Protein Intensity)
Proteins
LTQ XL Stem (4.7)
Mature Leaf (7.2)
Young Leaf (5.8)
Root (6.3)
Stem (6.0)

4.4 Conclusions

Since the release of the *Populus* genome in 2006, a question remains unanswered, i.e., what is the achievable depth and coverage of the predicted proteome space of Populus using high-throughput mass spectrometry? Although the application of bottom-up proteomics to measure global molecular responses is successful for many proteomic samples of low complexity, the depth of coverage required for a similar inquiry in higher eukaryotes requires more sophisticated sample preparation and advanced instrumentation. Therefore, we sought to address these issues by implementing a myriad of optimizations for nearly every step of the experimental process. These optimizations, while beneficial to plant proteomics in general, are broadly applicable to other organisms of similar complexity such as humans and other higher eukaryotes.

The enormous biological dynamic range inherent to a eukaryotic system demanded incorporation of a detergent-based sample preparation strategy that enhances plant cell lysis and protein extraction, both crucial enablers for in-depth analyses of complex proteomes. Without the appropriate instrumentation, this complexity inevitably leads to a sub-optimal identification of all detectable peptide species. Although a longstanding general challenge in shotgun proteomic experiments, recent technological improvements to the LTQ platform, mainly through enhancements to sequencing speed and sensitivity, doubles the identification rate of these dense peptide populations and enhances the identification of low-abundant protein species. Taken together, the enhanced sample preparation method and the state-of-the-art instrumentation enabled us to achieve one of the deepest proteome analyses in plant organisms to date, spanning six orders of magnitude in protein abundance and requiring only modest levels of sampling (i.e., 5-6 sample replicates per organ).

As demonstrated, the depth of coverage achieved in this study enables a comprehensive characterization of different plant organs that, at the cellular level, have vastly different chemical backgrounds (expressed genes, proteins and metabolites). As discussed in Chapter 5, this streamlined approach applied here affords an unprecedented view of *Populus* protein expression across several major organ-types.

CHAPTER 5

PUTTING THE PIECES TOGETHER: HIGH-PERFORMANCE LC-MS/MS PROVIDES NETWORK-, PATHWAY-, AND PROTEIN-LEVEL PERSPECTIVES IN *POPULUS*

All of the data presented below has been adapted from the following published journal article:

Paul Abraham, Richard Giannone, Rachel Adams, Udaya Kalluri, Gerald Tuskan, Robert Hettich (2012). "Putting the Pieces Together: High-performance LC-MS/MS Provides Network-, Pathway-, and Protein-level Perspectives in Populus". *Molecular and Cellular Proteomics 2013* 12(1): 106-119. Sample preparation and mass spectrometry experiments were performed by Paul Abraham. Biological data analysis was performed by Paul Abraham and Richard Giannone.

5.1 Introduction to the *Populus* Proteome Atlas

For most terrestrial plants, life begins and ends in the same physical location. For woody perennial plants, this sedentary lifestyle may last thousands of years. One consequence of this lifestyle is that each plant typically experiences dramatic changes in its ambient environment throughout its lifetime and, at any given time, equilibrium between endogenous growth processes and exogenous constraints exerted by the environment must be tightly controlled. To survive under varying environmental conditions, temporal plastic responses evoke patterns of protein expression that progressively influence morphological, anatomical and functional traits of three principal organs -- leaf, root and stem. Collectively and individually, these organs operate to perceive and respond to periodic and chronic environment conditions. Currently, a comprehensive understanding of the spatial variation in protein expression patterns across the organ types is lacking for woody perennial plants, where most large-scale proteome analyses with *Populus* were performed on isolated organs, tissues, organelles, or subcellular structures. For this reason, we combined the state-of-the-art LTQ-Velos platform with the SDS/TCA sample preparation methodology to generate a high-coverage proteome atlas of the principal organ types from *Populus* (see Chapter 4). With a high-coverage proteome atlas of the principal organ types, we provide a detailed look

into the predicted proteome space of *Populus*, offering varying proteome perspectives: 1) network-wide, 2) pathway-specific, and 3) protein-level viewpoints. As a demonstration of the precision and comprehensiveness of analysis, we also contrast two stages of leaf development, mature versus young leaf.

5.2 Profiling Organ-Specific Proteomes

5.2.1 Spatial Proteomics

Using the filtered and normalized data collected in Chapter 4, a function-level view of the different *Populus* organ proteomes was generated by sorting protein groups into functional categories as defined in the eukaryotic clusters of orthologous groups (KOG) database and weighting by normalized spectral count (nSpC) (Figure 5.1). KOG categories of "unknown function" and "post-translational modification and chaperone" had the highest representation in all organs. With regard to specific organs, "signal transduction mechanisms" and "chloroplast components" were the most abundant functional categories in mature leaves, "translation and RNA processing" in young leaves, "cytoskeleton" in stem and "unknown function" in roots.

We next identified protein groups from our data set that overlapped different organs, as well as those that were only found in one organ (Figure 5.2A). In the current study, a "core" proteome shared among the four different *Populus* organs was identified, consisting of 2,060 protein groups. The spatial distance between organs appeared to influence the degree of overlap between the different proteomes. For two organs that have a distal relationship, such as root and mature leaf, the overlap between proteomes decreases.

Protein groups found in only one organ may be linked to specialized, organ-specific processes. In total, we identified 688 protein groups unique to root, 831 to stem, 370 to mature leaf, and 629 to young leaf. For a more detailed comparison of the different organ proteomes, a Pearson correlation matrix assessed the correlation between the different organ proteomes (Figure 5.2B).

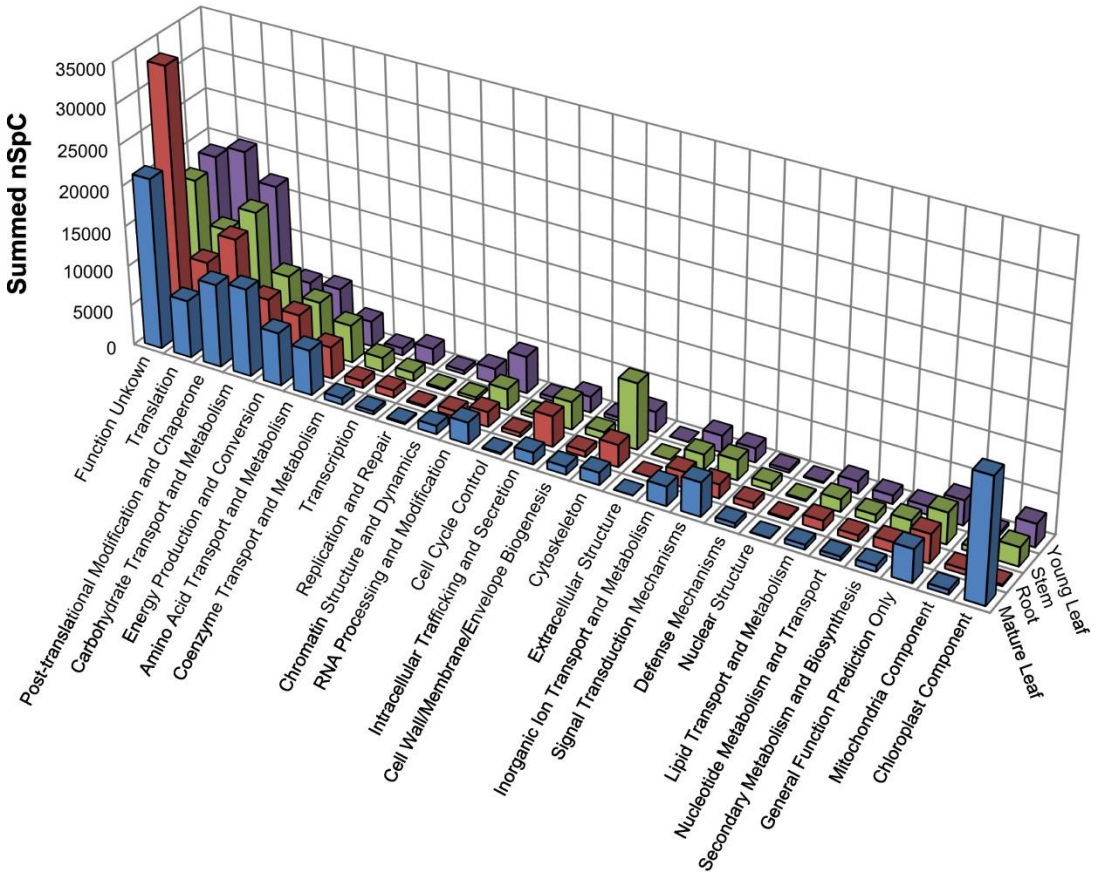


Figure 5.1. Quantitative distribution of detected protein groups by their KOG functional classification categories. Proteins identified in each tissue type were assigned KOG categories to identify functional trends relevant to each organ type. Category representation was weighted by the sum total of the normalized spectral counts (nSpC) contributed by each protein in the classification. Notable trends include a high proportion of nSpC in mature leaf attributed to chloroplast-based proteins, enrichment of cytoskeletal components in stem, and an increase in translation in young leaf compared to the other tissues. Also noted is the large degree of nSpC representation falling into the unknown category, suggesting a need for improved protein annotation as a whole.

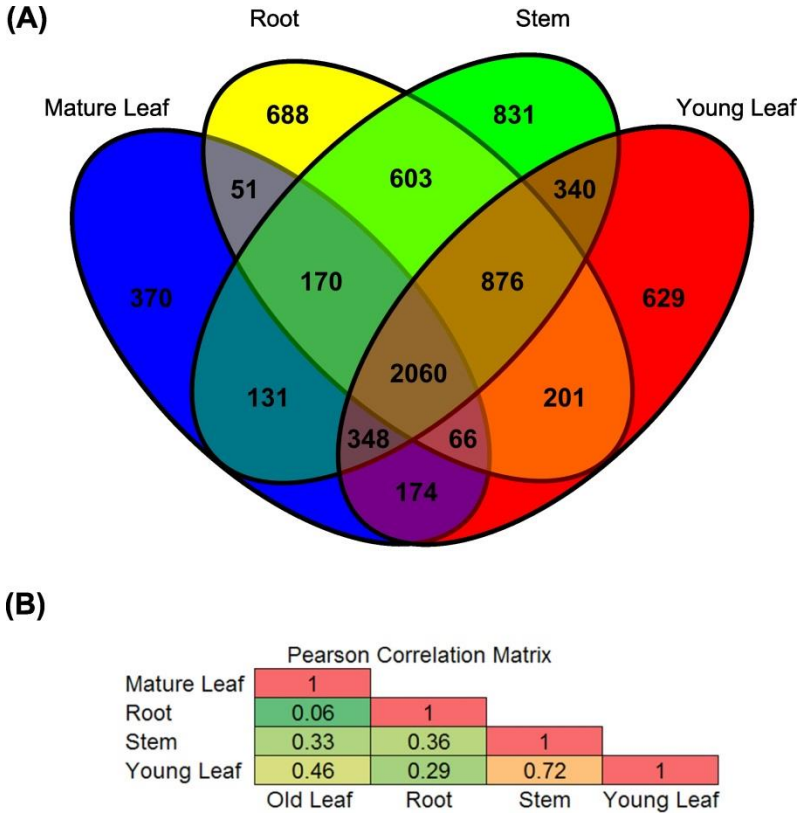


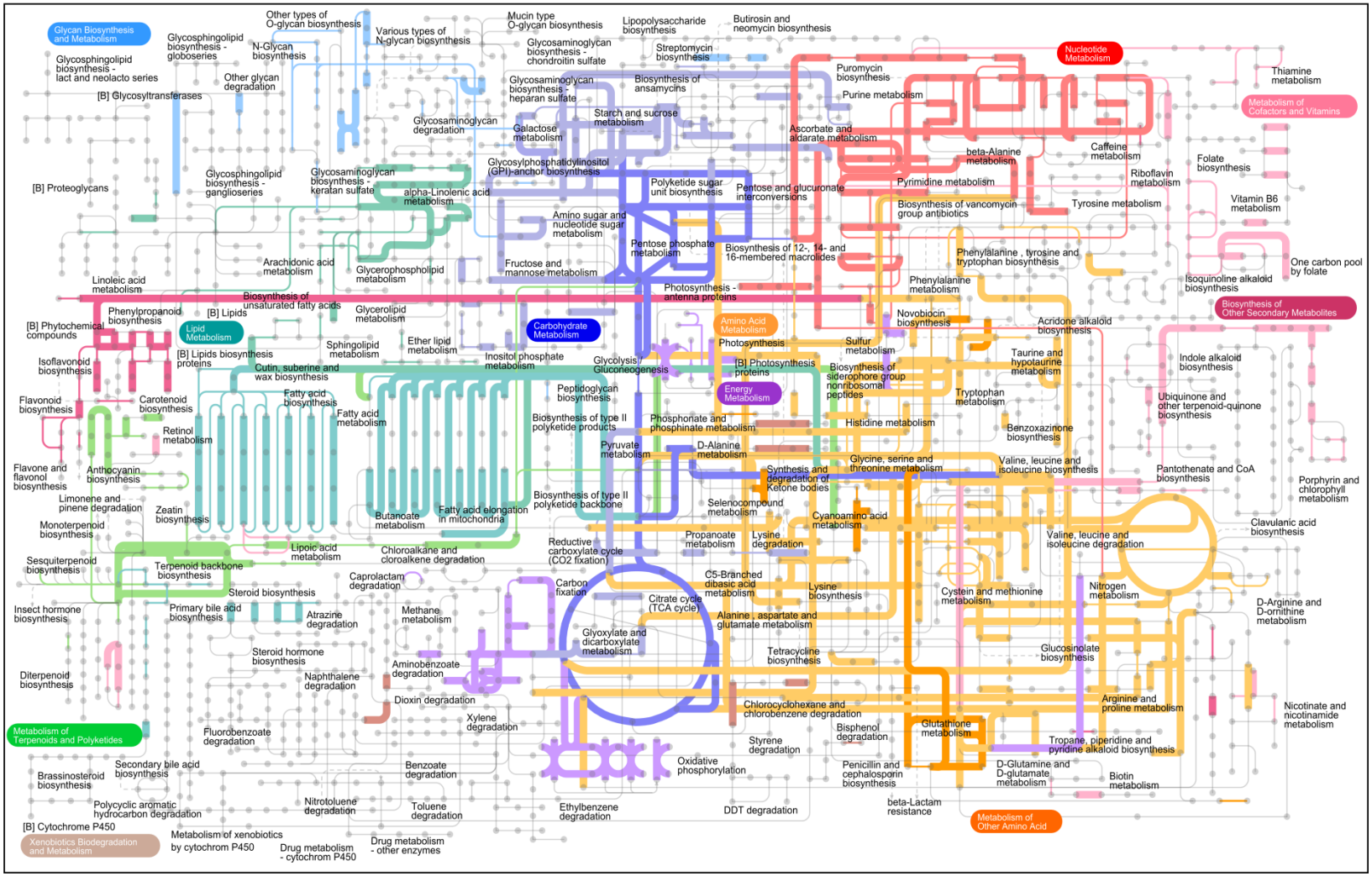
Figure 5.2. Global proteomic view across all four *Populus* organs. Numbers of identified protein groups, as represented by a 4-tiered Venn diagram (A), indicate the level of proteomic overlap between organ types. Notable regions include protein groups specific to only one organ-type (solid blue, yellow, green and red) as well as groups identified across all organs (central brown region). (B) Degree of proteomic overlap as visualized by Pearson's correlation analysis of all protein nSpC values averaged across all replicates for each particular organ. The degree of correlation increases as a function of organ proximity.

The pairwise comparisons resulted in Pearson correlation values that range from 0.06 (mature leaf vs. root) to 0.72 (young leaf vs. stem). The correlation coefficients support the results in Figure 5.2A and together corroborate the hypothesis that the degree of proteomic overlap between different *Populus* organs is reflected by their shared function.

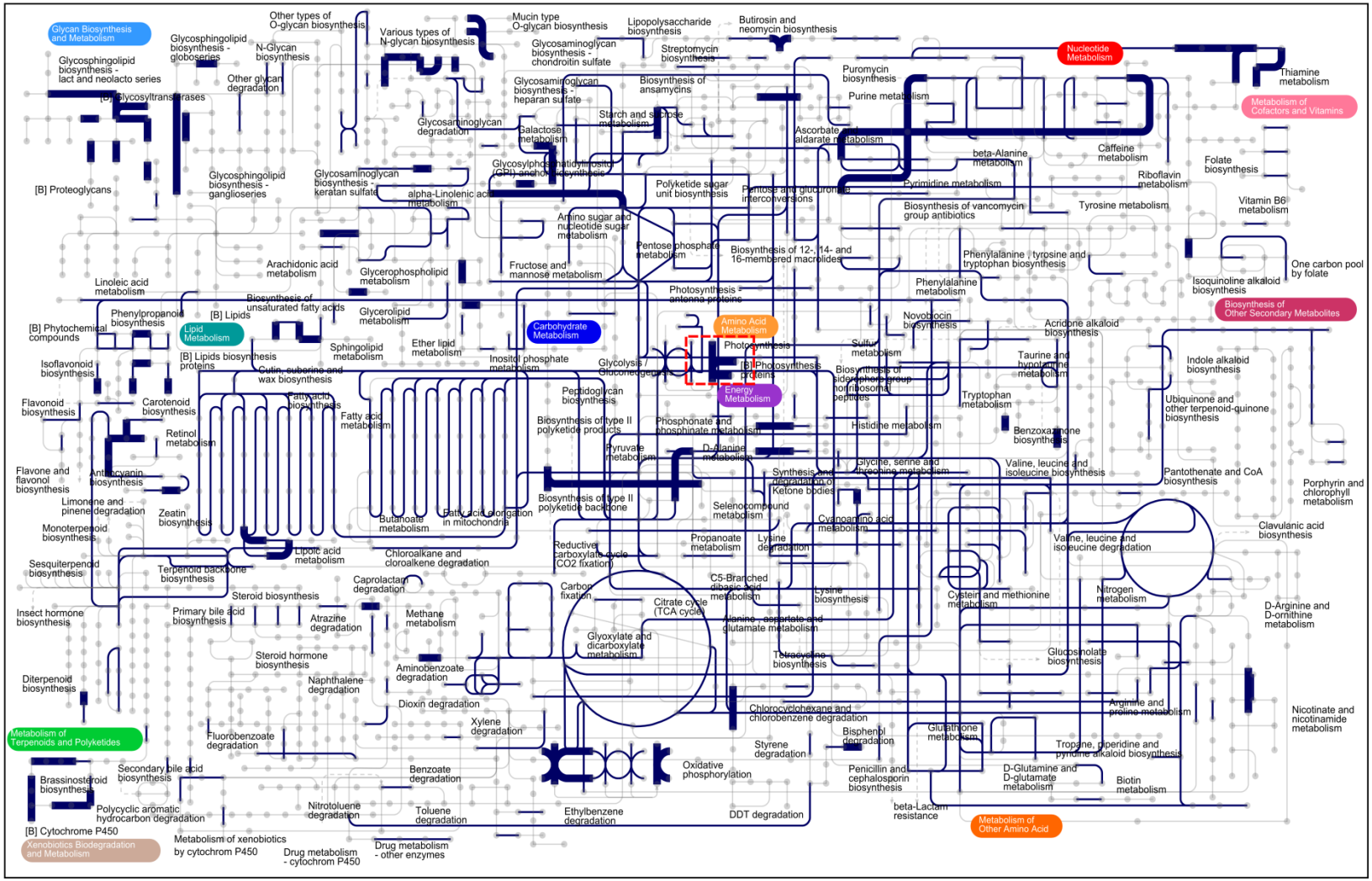
For a network-wide perspective of *Populus* metabolism, we employed the use of iPath2.0[230] (http://pathways.embl.de) to navigate and explore the predicted KEGG metabolic pathways (Figure 5.3A-E). Using the entire data set (7,538 protein groups), a metabolic pathway diagram was constructed to highlight the core proteome relative to all protein groups measured (Figure 5.3A). Overall, the core molecular network spanned every major functional category belonging to central metabolism. These protein groups likely belong to catalytic and regulatory interactions that govern the life of a plant cell, and may include signaling networks that choreograph cross talk between plant cells in response to environmental perturbations. Figures 5.3B-E depict metabolic grids of individual organ proteomes. Even though similar coverage of the metabolic network was observed for each organ, the most revealing feature of these maps is the existence of molecular networks and the protein groups that are characteristic of a specific organ. For each organ proteome, a number of unique protein groups were identified and, rather than mapping ubiquitously, they generally assembled into discrete pathways. Although beyond the scope of this study, future work could integrate metabolomics to measure net fluxes of material into and out of pathways, capturing the relationship of enzymes and their substrates/products[231].

Figure 5.3. Metabolic pathway maps for *Populus*. The data set was represented in a visual metabolic context by integrating identified protein groups with the predicted KEGG information for *P. trichocarpa*. Metabolic maps were created for the following proteomes: A) core, B) mature leaf, C) young leaf, D) root and E) stem (including both phloem and xylem tissues). For each panel, the grey nodes represent various chemical compounds and each line represents an enzymatic reaction. Lines that are grey represent enzyme reactions in which no protein group was identified. A colored line represents an enzymatic reaction in which a protein group was identified. Line widths in the map denote uniqueness of a protein group for a particular proteome. For example, a thick colored line in A represents a protein group that was found across all organ types and thus a member of the core proteome. To summarize the metabolic state of each organ, a few representative modules were chosen (outlined by a red dashed border). For mature leaf (B), a module was chosen to represent the increased number of proteins involved in photosynthesis. This module includes 7 subunits of the F-type H+-transporting ATPase enzyme found in chloroplast thylakoid membranes. A module of protein groups encircled in nucleotide metabolism was chosen to illustrate the increased rate of growth in young leaf (C). Roots (D) are critical storage hub in plants; consequently, a module of protein groups belonging to starch and sucrose metabolism were highlighted. For increased structural stability, stem (E) tissue consists of a milieu of protein groups belonging to lipid metabolism.

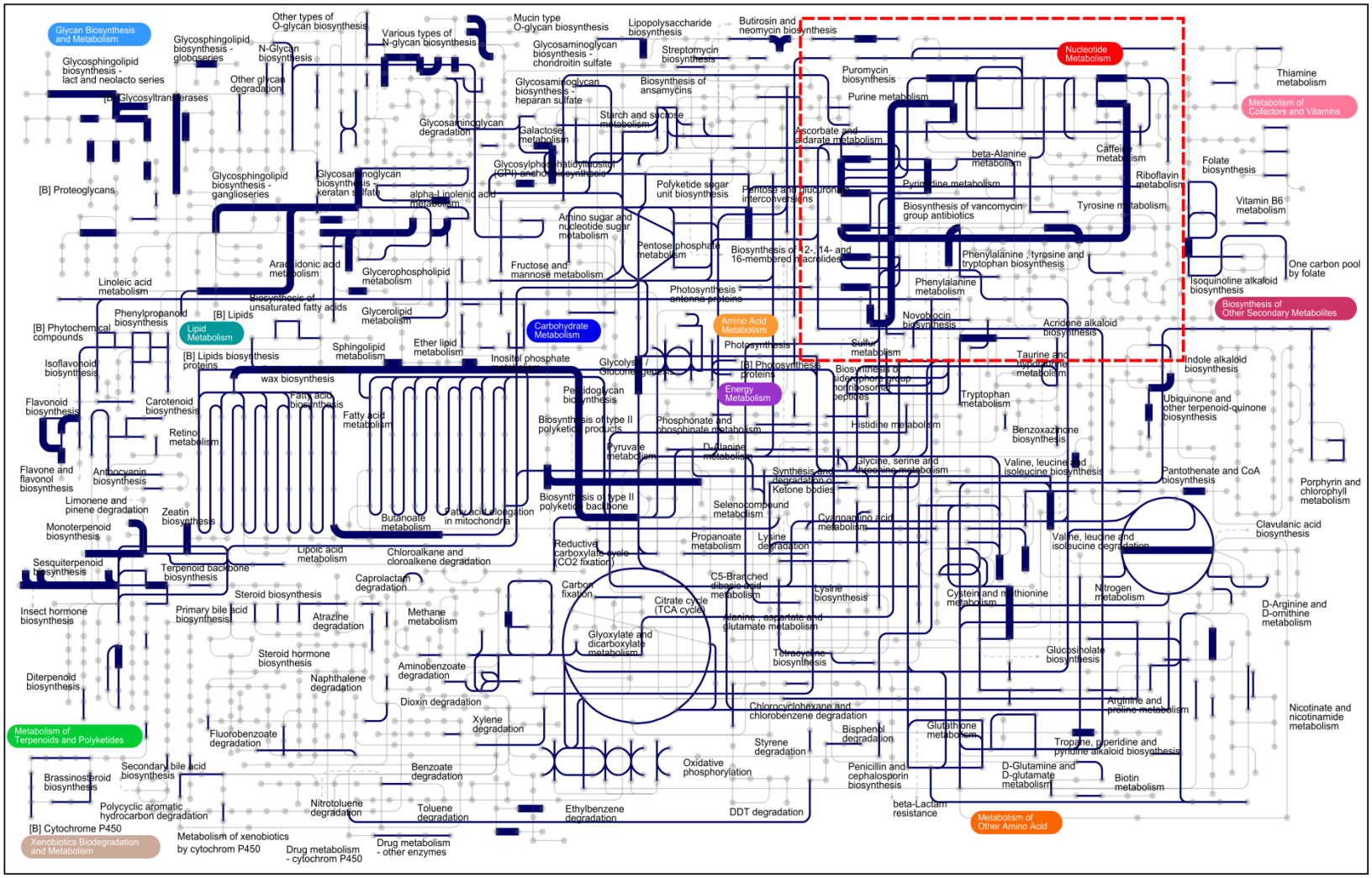
(A)


Glycan Biosynthesis and Metabolism
Glycosphingolipid biosynthesis - lact and neolacto series
Glycosphingolipid biosynthesis - globoseries
Other types of O-glycan biosynthesis
N-Glycan biosynthesis
Other glycan degradation
Various types of N-glycan biosynthesis
[B] Glycosyltransferases
[B] Proteoglycans
Glycosphingolipid biosynthesis - ganglioseries
Glycosaminoglycan biosynthesis - keratan sulfate
Glycosaminoglycan degradation
alpha-Linolenic acid metabolism
Mucin type O-glycan biosynthesis
Glycosaminoglycan biosynthesis - chondroitin sulfate
Glycosaminoglycan biosynthesis - heparan sulfate
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis
Lipopolysaccharide biosynthesis
Butirosin and neomycin biosynthesis
Streptomycin biosynthesis
Biosynthesis of ansamycins
Starch and sucrose metabolism
Galactose metabolism
Amino sugar and nucleotide sugar metabolism
Fructose and mannose metabolism
Pentose phosphate metabolism
Polyketide sugar unit biosynthesis
Pentose and glucuronate interconversions
Biosynthesis of 12-, 14- and 16-membered macrolides
Arachidonic acid metabolism
Glycerophospholipid metabolism
Linoleic acid metabolism
Biosynthesis of unsaturated fatty acids
Glycerolipid metabolism
Photosynthesis - antenna proteins
[B] Phytochemical compounds
Phenylpropanoid biosynthesis
[B] Lipids
Lipid Metabolism
[B] Lipids biosynthesis proteins
Sphingolipid metabolism
Ether lipid metabolism
Inositol phosphate metabolism
Cutin, suberine and wax biosynthesis
Isoflavonoid biosynthesis
Flavonoid biosynthesis
Carotenoid biosynthesis
Retinol metabolism
Fatty acid biosynthesis
Fatty acid metabolism
Carbohydrate Metabolism
Glycolysis / Gluconeogenesis
Peptidoglycan biosynthesis
Biosynthesis of type II polyketide products
Pyruvate metabolism
Amino Acid Metabolism
Photosynthesis
[B] Photosynthesis proteins
Energy Metabolism
Phosphonate and phosphinate metabolism
D-Alanine metabolism
Sulfur metabolism
Biosynthesis of siderophore group nonribosomal peptides
Histidine metabolism
Flavone and flavonol biosynthesis
Anthocyanin biosynthesis
Limonene and pinene degradation
Zeatin biosynthesis
Monoterpenoid biosynthesis
Sesquiterpenoid biosynthesis
Terpenoid backbone biosynthesis
Steroid biosynthesis
Insect hormone biosynthesis
Primary bile acid biosynthesis
Diterpenoid biosynthesis
Metabolism of Terpenoids and Polyketides
Secondary bile acid biosynthesis
Brassinosteroid biosynthesis
Biosynthesis of type II polyketide backbone
Butanoate metabolism
Fatty acid elongation in mitochondria
Lipoic acid metabolism
Chloroalkane and chloroalkene degradation
Caprolactam degradation
Atrazine degradation
Methane metabolism
Steroid hormone biosynthesis
Naphthalene degradation
Aminobenzoate degradation
Dioxin degradation
Fluorobenzoate degradation
Xylene degradation
Benzoate degradation
Polycyclic aromatic hydrocarbon degradation
Nitrotoluene degradation
Toluene degradation
Ethylbenzene degradation
[B] Cytochrome P450
Xenobiotics Biodegradation and Metabolism
Metabolism of xenobiotics by cytochrom P450
Drug metabolism - cytochrom P450
Drug metabolism - other enzymes
Selenocompound metabolism
Propanoate metabolism
Lysine degradation
Reductive carboxylate cycle (CO2 fixation)
Carbon fixation
Citrate cycle (TCA cycle)
C5-Branched dibasic acid metabolism
Alanine , aspartate and glutamate metabolism
Glyoxylate and dicarboxylate metabolism
Synthesis and degradation of Ketone bodies
Glycine, serine and threonine metabolism
Cyanoamino acid metabolism
Lysine biosynthesis
Tetracycline biosynthesis
Chlorocyclohexane and chlorobenzene degradation
Styrene degradation
Oxidative phosphorylation
DDT degradation
Nucleotide Metabolism
Puromycin biosynthesis
Purine metabolism
Ascorbate and aldarate metabolism
beta-Alanine metabolism
Pyrimidine metabolism
Biosynthesis of vancomycin group antibiotics
Phenylalanine , tyrosine and tryptophan biosynthesis
Phenylalanine metabolism
Novobiocin biosynthesis
Photosynthesis
Sulfur metabolism
Taurine and hypotaurine metabolism
Tryptophan metabolism
Histidine metabolism
Benzoxazinone biosynthesis
Valine, leucine and isoleucine biosynthesis
Valine, leucine and isoleucine degradation
Cysteine and methionine metabolism
Glucosinolate biosynthesis
Arginine and proline metabolism
Glutathione metabolism
Bisphenol degradation
Penicillin and cephalosporin biosynthesis
Tropane, piperidine and pyridine alkaloid biosynthesis
D-Glutamine and D-glutamate metabolism
Biotin metabolism
beta-Lactam resistance
Metabolism of Other Amino Acid
Nitrogen metabolism
Caffeine metabolism
Riboflavin metabolism
Tyrosine metabolism
Thiamine metabolism
Metabolism of Cofactors and Vitamins
Folate biosynthesis
Vitamin B6 metabolism
One carbon pool by folate
Isoquinoline alkaloid biosynthesis
Biosynthesis of Other Secondary Metabolites
Acridone alkaloid biosynthesis
Indole alkaloid biosynthesis
Ubiquinone and other terpenoid-quinone biosynthesis
Pantothenate and CoA biosynthesis
Porphyrin and chlorophyll metabolism
Clavulanic acid biosynthesis
D-Arginine and D-ornithine metabolism
Nicotinate and nicotinamide metabolism

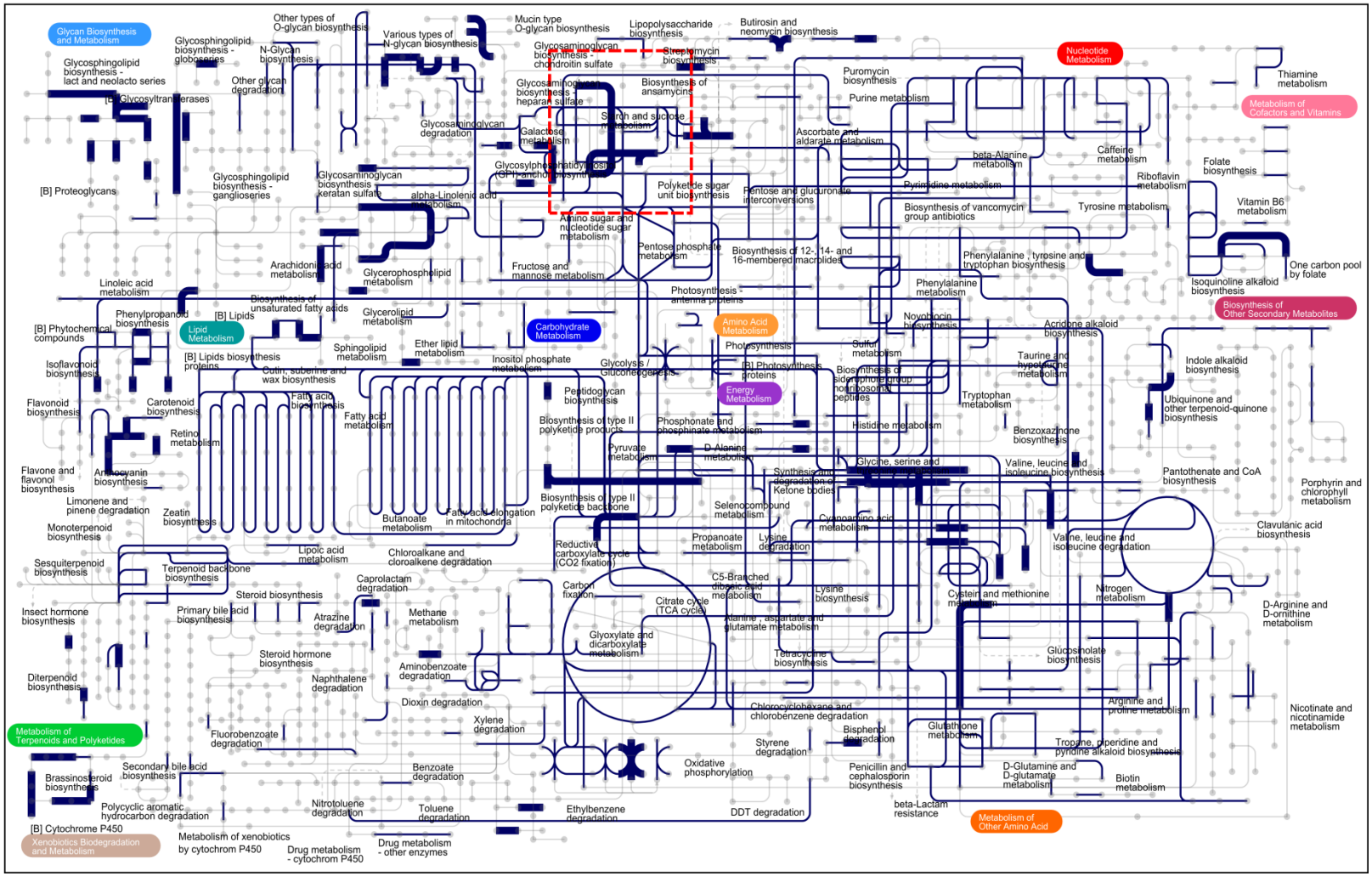
(B)


Glycan Biosynthesis and Metabolism
Glycosphingolipid biosynthesis - lact and neolacto series
[B] Glycosyltransferases
[B] Proteoglycans
Glycosphingolipid biosynthesis - globoseries
Other types of O-glycan biosynthesis
N-Glycan biosynthesis
Other glycan degradation
Various types of N-glycan biosynthesis
Glycosphingolipid biosynthesis - ganglioseries
Glycosaminoglycan biosynthesis - keratan sulfate
Glycosaminoglycan degradation
alpha-Linolenic acid metabolism
Mucin type O-glycan biosynthesis
Glycosaminoglycan biosynthesis - chondroitin sulfate
Glycosaminoglycan biosynthesis - heparan sulfate
Lipopolysaccharide biosynthesis
Streptomycin biosynthesis
Butirosin and neomycin biosynthesis
Biosynthesis of ansamycins
Starch and sucrose metabolism
Galactose metabolism
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis
Amino sugar and nucleotide sugar metabolism
Polyketide sugar unit biosynthesis
Pentose and glucuronate interconversions
Pentose phosphate metabolism
Fructose and mannose metabolism
Biosynthesis of 12-, 14- and 16-membered macrolides
Photosynthesis - antenna proteins
Arachidonic acid metabolism
Glycerophospholipid metabolism
Linoleic acid metabolism
Biosynthesis of unsaturated fatty acids
Glycerolipid metabolism
Phenylpropanoid biosynthesis
[B] Lipids
[B] Phytochemical compounds
Lipid Metabolism
[B] Lipids biosynthesis proteins
Sphingolipid metabolism
Ether lipid metabolism
Cutin, suberine and wax biosynthesis
Isoflavonoid biosynthesis
Flavonoid biosynthesis
Carotenoid biosynthesis
Fatty acid biosynthesis
Fatty acid metabolism
Retinol metabolism
Flavone and flavonol biosynthesis
Anthocyanin biosynthesis
Limonene and pinene degradation
Zeatin biosynthesis
Monoterpenoid biosynthesis
Sesquiterpenoid biosynthesis
Terpenoid backbone biosynthesis
Steroid biosynthesis
Insect hormone biosynthesis
Primary bile acid biosynthesis
Steroid hormone biosynthesis
Diterpenoid biosynthesis
Metabolism of Terpenoids and Polyketides
Brassinosteroid biosynthesis
Secondary bile acid biosynthesis
Polycyclic aromatic hydrocarbon degradation
[B] Cytochrome P450
Xenobiotics Biodegradation and Metabolism
Metabolism of xenobiotics by cytochrom P450
Drug metabolism - cytochrom P450
Drug metabolism - other enzymes
Carbohydrate Metabolism
Glycolysis / Gluconeogenesis
Inositol phosphate metabolism
Peptidoglycan biosynthesis
Biosynthesis of type II polyketide products
Biosynthesis of type II polyketide backbone
Pyruvate metabolism
Butanoate metabolism
Fatty acid elongation in mitochondria
Lipoic acid metabolism
Chloroalkane and chloroalkene degradation
Caprolactam degradation
Carbon fixation
Reductive carboxylate cycle (CO2 fixation)
Citrate cycle (TCA cycle)
Glyoxylate and dicarboxylate metabolism
Propanoate metabolism
Lysine degradation
C5-Branched dibasic acid metabolism
Alanine, aspartate and glutamate metabolism
Methane metabolism
Atrazine degradation
Aminobenzoate degradation
Naphthalene degradation
Dioxin degradation
Xylene degradation
Fluorobenzoate degradation
Benzoate degradation
Nitrotoluene degradation
Toluene degradation
Ethylbenzene degradation
Oxidative phosphorylation
Amino Acid Metabolism
Photosynthesis
[B] Photosynthesis proteins
Energy Metabolism
Phosphonate and phosphinate metabolism
D-Alanine metabolism
Synthesis and degradation of Ketone bodies
Selenocompound metabolism
Cyanoamino acid metabolism
Lysine biosynthesis
Tetracycline biosynthesis
Chlorocyclohexane and chlorobenzene degradation
Styrene degradation
DDT degradation
Nucleotide Metabolism
Puromycin biosynthesis
Purine metabolism
Ascorbate and aldarate metabolism
beta-Alanine metabolism
Pyrimidine metabolism
Biosynthesis of vancomycin group antibiotics
Phenylalanine, tyrosine and tryptophan biosynthesis
Phenylalanine metabolism
Novobiocin biosynthesis
Sulfur metabolism
Biosynthesis of siderophore group nonribosomal peptides
Histidine metabolism
Glycine, serine and threonine metabolism
Valine, leucine and isoleucine degradation
Cysteine and methionine metabolism
Penicillin and cephalosporin biosynthesis
Glutathione metabolism
Bisphenol degradation
beta-Lactam resistance
Metabolism of Other Amino Acid
Thiamine metabolism
Metabolism of Cofactors and Vitamins
Caffeine metabolism
Folate biosynthesis
Riboflavin metabolism
Vitamin B6 metabolism
Tyrosine metabolism
One carbon pool by folate
Isoquinoline alkaloid biosynthesis
Biosynthesis of Other Secondary Metabolites
Acridone alkaloid biosynthesis
Taurine and hypotaurine metabolism
Tryptophan metabolism
Benzoxazinone biosynthesis
Valine, leucine and isoleucine biosynthesis
Indole alkaloid biosynthesis
Ubiquinone and other terpenoid-quinone biosynthesis
Pantothenate and CoA biosynthesis
Porphyrin and chlorophyll metabolism
Clavulanic acid biosynthesis
Nitrogen metabolism
D-Arginine and D-ornithine metabolism
Glucosinolate biosynthesis
Arginine and proline metabolism
Nicotinate and nicotinamide metabolism
Tropane, piperidine and pyridine alkaloid biosynthesis
D-Glutamine and D-glutamate metabolism
Biotin metabolism

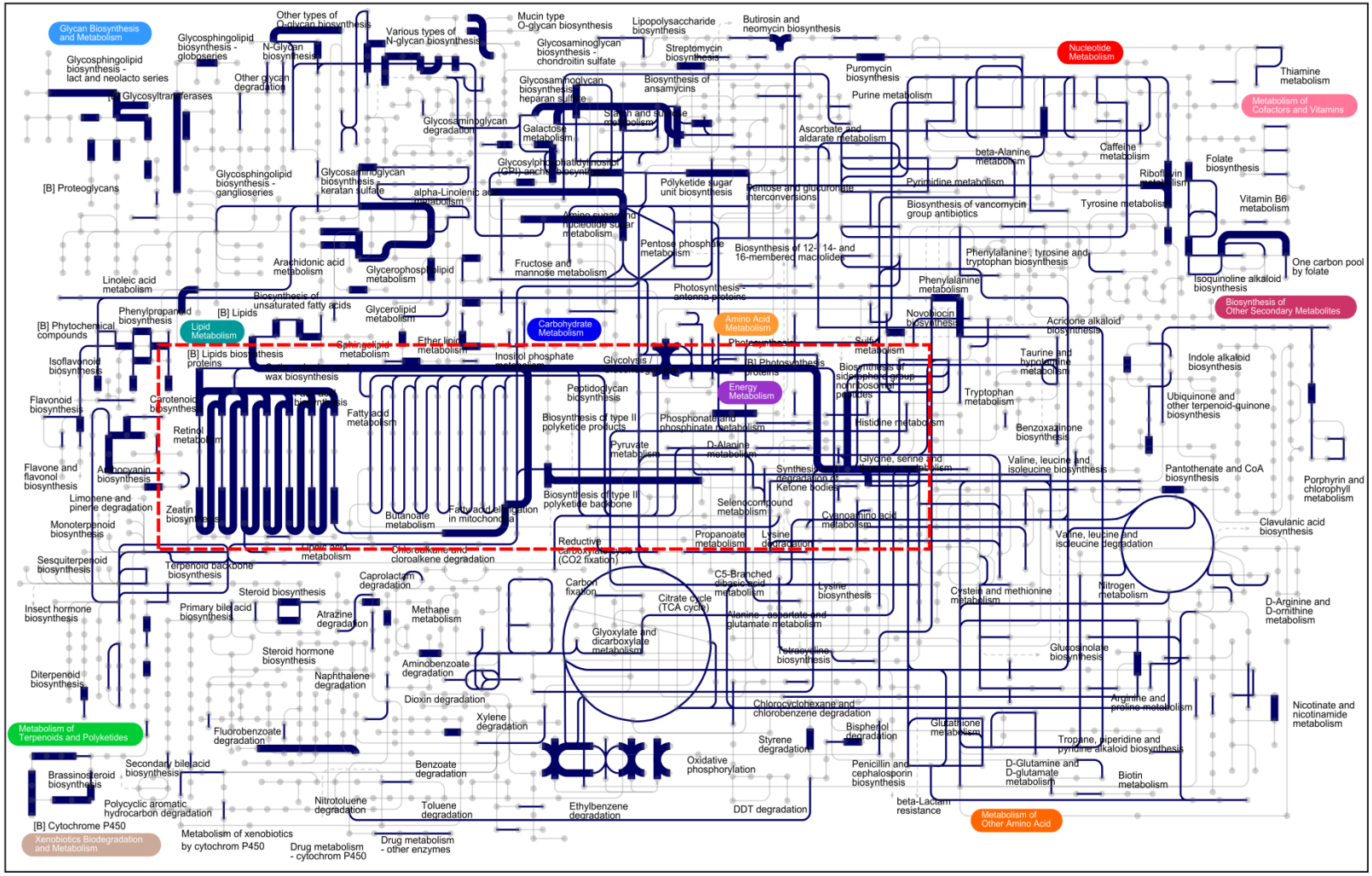
(C)


Glycan Biosynthesis and Metabolism
Glycosphingolipid biosynthesis - lacto and neolacto series
Glycosphingolipid biosynthesis - globoseries
Other types of O-glycan biosynthesis
N-Glycan biosynthesis
Various types of N-glycan biosynthesis
Other glycan degradation
Glycosyltransferases
[B] Proteoglycans
Glycosphingolipid biosynthesis - ganglioseries
Glycosaminoglycan biosynthesis - keratan sulfate
Glycosaminoglycan degradation
Mucin type O-glycan biosynthesis
Glycosaminoglycan biosynthesis - chondroitin sulfate
Glycosaminoglycan biosynthesis - heparan sulfate
Lipopolysaccharide biosynthesis
Streptomycin biosynthesis
Butirosin and neomycin biosynthesis
Biosynthesis of ansamycins
Starch and sucrose metabolism
Galactose metabolism
Glycosylphosphatidylinositol (GPI) anchor biosynthesis
alpha-Linolenic acid metabolism
Polyketide sugar unit biosynthesis
Pentose and glucuronate interconversions
Amino sugar and nucleotide sugar metabolism
Pentose phosphate pathway
Biosynthesis of 12-, 14- and 16-membered macrolides
Fructose and mannose metabolism
Arachidonic acid metabolism
Glycerophospholipid metabolism
Linoleic acid metabolism
Biosynthesis of unsaturated fatty acids
Glycerolipid metabolism
Photosynthesis - antenna proteins
Phenylpropanoid biosynthesis
[B] Lipids
Lipid Metabolism
[B] Phytochemical compounds
[B] Lipids biosynthesis proteins
Sphingolipid metabolism
Ether lipid metabolism
Inositol phosphate metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Photosynthesis
[B] Photosynthesis proteins
Energy Metabolism
Glycolysis / Gluconeogenesis
Peptidoglycan biosynthesis
Biosynthesis of type II polyketide products
Phosphonate and phosphinate metabolism
Pyruvate metabolism
D-Alanine metabolism
Fatty acid biosynthesis
Fatty acid metabolism
Isoflavonoid biosynthesis
Flavonoid biosynthesis
Carotenoid biosynthesis
Retinol metabolism
Flavone and flavonol biosynthesis
Anthocyanin biosynthesis
Limonene and pinene degradation
Zeatin biosynthesis
Monoterpenoid biosynthesis
Sesquiterpenoid biosynthesis
Terpenoid backbone biosynthesis
Steroid biosynthesis
Insect hormone biosynthesis
Primary bile acid biosynthesis
Steroid hormone biosynthesis
Diterpenoid biosynthesis
Brassinosteroid biosynthesis
Secondary bile acid biosynthesis
Polycyclic aromatic hydrocarbon degradation
[B] Cytochrome P450
Xenobiotics Biodegradation and Metabolism
Metabolism of xenobiotics by cytochrom P450
Drug metabolism - cytochrom P450
Drug metabolism - other enzymes
Butanoate metabolism
Fatty acid elongation in mitochondria
Lipoic acid metabolism
Chloroalkane and chloroalkene degradation
Caprolactam degradation
Atrazine degradation
Methane metabolism
Naphthalene degradation
Aminobenzoate degradation
Dioxin degradation
Xylene degradation
Fluorobenzoate degradation
Benzoate degradation
Nitrotoluene degradation
Toluene degradation
Ethylbenzene degradation
Biosynthesis of type II polyketide backbone
Reductive carboxylate cycle (CO2 fixation)
Carbon fixation
Citrate cycle (TCA cycle)
Glyoxylate and dicarboxylate metabolism
Oxidative phosphorylation
Metabolism of Terpenoids and Polyketides
Nucleotide Metabolism
Puromycin biosynthesis
Purine metabolism
Ascorbate and aldarate metabolism
beta-Alanine metabolism
Pyrimidine metabolism
Biosynthesis of vancomycin group antibiotics
Caffeine metabolism
Tyrosine metabolism
Riboflavin metabolism
Thiamine metabolism
Metabolism of Cofactors and Vitamins
Folate biosynthesis
Vitamin B6 metabolism
Phenylalanine, tyrosine and tryptophan biosynthesis
Phenylalanine metabolism
Novobiocin biosynthesis
Sulfur metabolism
Taurine and hypotaurine metabolism
Acridone alkaloid biosynthesis
Isoquinoline alkaloid biosynthesis
One carbon pool by folate
Biosynthesis of Other Secondary Metabolites
Indole alkaloid biosynthesis
Ubiquinone and other terpenoid-quinone biosynthesis
Tryptophan metabolism
Histidine metabolism
Benzoxazinone biosynthesis
Pantothenate and CoA biosynthesis
Porphyrin and chlorophyll metabolism
Valine, leucine and isoleucine biosynthesis
Glycine, serine and threonine metabolism
Synthesis and degradation of ketone bodies
Selenocompound metabolism
Cysteine and methionine metabolism
Lysine biosynthesis
Lysine degradation
Propanoate metabolism
Alanine, aspartate and glutamate metabolism
C5-Branched dibasic acid metabolism
Tetracycline biosynthesis
Valine, leucine and isoleucine degradation
Clavulanic acid biosynthesis
Nitrogen metabolism
D-Arginine and D-ornithine metabolism
Glucosinolate biosynthesis
Arginine and proline metabolism
Nicotinate and nicotinamide metabolism
Chlorocyclohexane and chlorobenzene degradation
Bisphenol degradation
Styrene degradation
Glutathione metabolism
Tropane, piperidine and pyridine alkaloid biosynthesis
Penicillin and cephalosporin biosynthesis
D-Glutamine and D-glutamate metabolism
Biotin metabolism
beta-Lactam resistance
DDT degradation
Metabolism of Other Amino Acid

(D)


Glycan Biosynthesis and Metabolism
Glycosphingolipid biosynthesis - globoseries
Other types of O-glycan biosynthesis
N-Glycan biosynthesis
Various types of N-glycan biosynthesis
Glycosphingolipid biosynthesis - lact and neolacto series
Other glycan degradation
[B] Glycosyltransferases
Mucin type O-glycan biosynthesis
Lipopolysaccharide biosynthesis
Butirosin and neomycin biosynthesis
Glycosaminoglycan biosynthesis - chondroitin sulfate
Streptomycin biosynthesis
Glycosaminoglycan biosynthesis - heparan sulfate
Biosynthesis of ansamycins
Glycosaminoglycan degradation
Starch and sucrose metabolism
Galactose metabolism
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis
Glycosphingolipid biosynthesis - ganglioseries
Glycosaminoglycan biosynthesis - keratan sulfate
[B] Proteoglycans
alpha-Linolenic acid metabolism
Polyketide sugar unit biosynthesis
Pentose and glucuronate interconversions
Amino sugar and nucleotide sugar metabolism
Pentose phosphate metabolism
Biosynthesis of 12-, 14- and 16-membered macrolides
Arachidonic acid metabolism
Glycerophospholipid metabolism
Fructose and mannose metabolism
Linoleic acid metabolism
Biosynthesis of unsaturated fatty acids
Glycerolipid metabolism
Photosynthesis - antenna proteins
Phenylpropanoid biosynthesis
[B] Lipids
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
[B] Phytochemical compounds
[B] Lipids biosynthesis proteins
Sphingolipid metabolism
Ether lipid metabolism
Inositol phosphate metabolism
Photosynthesis
Sulfur metabolism
Isoflavonoid biosynthesis
Cutin, suberine and wax biosynthesis
Glycolysis / Gluconeogenesis
[B] Photosynthesis proteins
Energy Metabolism
Pentoglycan biosynthesis
Flavonoid biosynthesis
Carotenoid biosynthesis
Fatty acid biosynthesis
Fatty acid metabolism
Biosynthesis of type II polyketide products
Phosphonate and phosphinate metabolism
Retinol metabolism
Pyruvate metabolism
D-Alanine metabolism
Flavone and flavonol biosynthesis
Anthocyanin biosynthesis
Synthesis and degradation of Ketone bodies
Limonene and pinene degradation
Zeatin biosynthesis
Biosynthesis of type II polyketide backbone
Selenocompound metabolism
Monoterpenoid biosynthesis
Butanoate metabolism
Fatty acid elongation in mitochondria
Cyanoamino acid metabolism
Lipoic acid metabolism
Chloroalkane and chloroalkene degradation
Propanoate metabolism
Lysine degradation
Sesquiterpenoid biosynthesis
Terpenoid backbone biosynthesis
Reductive carboxylate cycle (CO2 fixation)
Caprolactam degradation
C5-Branched dibasic acid metabolism
Carbon fixation
Steroid biosynthesis
Lysine biosynthesis
Insect hormone biosynthesis
Primary bile acid biosynthesis
Atrazine degradation
Methane metabolism
Citrate cycle (TCA cycle)
Alanine, aspartate and glutamate metabolism
Glyoxylate and dicarboxylate metabolism
Steroid hormone biosynthesis
Tetracycline biosynthesis
Diterpenoid biosynthesis
Naphthalene degradation
Aminobenzoate degradation
Dioxin degradation
Chlorocyclohexane and chlorobenzene degradation
Metabolism of Terpenoids and Polyketides
Xylene degradation
Fluorobenzoate degradation
Styrene degradation
Bisphenol degradation
Glutathione metabolism
Secondary bile acid biosynthesis
Benzoate degradation
Oxidative phosphorylation
Penicillin and cephalosporin biosynthesis
Brassinosteroid biosynthesis
Polycyclic aromatic hydrocarbon degradation
Nitrotoluene degradation
Toluene degradation
Ethylbenzene degradation
DDT degradation
beta-Lactam resistance
[B] Cytochrome P450
Xenobiotics Biodegradation and Metabolism
Metabolism of xenobiotics by cytochrom P450
Drug metabolism - cytochrom P450
Drug metabolism - other enzymes
Nucleotide Metabolism
Puromycin biosynthesis
Purine metabolism
Thiamine metabolism
Metabolism of Cofactors and Vitamins
Ascorbate and aldarate metabolism
beta-Alanine metabolism
Caffeine metabolism
Folate biosynthesis
Pyrimidine metabolism
Riboflavin metabolism
Biosynthesis of vancomycin group antibiotics
Tyrosine metabolism
Vitamin B6 metabolism
Phenylalanine, tyrosine and tryptophan biosynthesis
One carbon pool by folate
Phenylalanine metabolism
Isoquinoline alkaloid biosynthesis
Novobiocin biosynthesis
Biosynthesis of Other Secondary Metabolites
Acridone alkaloid biosynthesis
Taurine and hypotaurine metabolism
Indole alkaloid biosynthesis
Biosynthesis of siderophore group nonribosomal peptides
Tryptophan metabolism
Ubiquinone and other terpenoid-quinone biosynthesis
Histidine metabolism
Benzoxazinone biosynthesis
Valine, leucine and isoleucine biosynthesis
Pantothenate and CoA biosynthesis
Porphyrin and chlorophyll metabolism
Glycine, serine and
Valine, leucine and isoleucine degradation
Clavulanic acid biosynthesis
Cysteine and methionine metabolism
Nitrogen metabolism
D-Arginine and D-ornithine metabolism
Glucosinolate biosynthesis
Arginine and proline metabolism
Nicotinate and nicotinamide metabolism
Tropane, piperidine and pyridine alkaloid biosynthesis
D-Glutamine and D-glutamate metabolism
Biotin metabolism
Metabolism of Other Amino Acid

(E)


Glycan Biosynthesis and Metabolism
Glycosphingolipid biosynthesis - lact and neolacto series
Glycosphingolipid biosynthesis - globoseries
Other types of O-glycan biosynthesis
N-Glycan biosynthesis
Other glycan degradation
Various types of N-glycan biosynthesis
Mucin type O-glycan biosynthesis
Glycosaminoglycan biosynthesis - chondroitin sulfate
Lipopolysaccharide biosynthesis
Streptomycin biosynthesis
Butirosin and neomycin biosynthesis
Glycosaminoglycan biosynthesis - heparan sulfate
Biosynthesis of ansamycins
Glycosaminoglycan degradation
Starch and sucrose metabolism
Galactose metabolism
Ascorbate and aldarate metabolism
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis
Glycosphingolipid biosynthesis - ganglioseries
Glycosaminoglycan biosynthesis - keratan sulfate
alpha-Linolenic acid metabolism
Polyketide sugar unit biosynthesis
Pentose and glucuronate interconversions
[B] Proteoglycans
Amino sugar and nucleotide sugar metabolism
Pentose phosphate metabolism
Biosynthesis of 12-, 14- and 16-membered macrolides
Arachidonic acid metabolism
Glycerophospholipid metabolism
Fructose and mannose metabolism
Linoleic acid metabolism
Biosynthesis of unsaturated fatty acids
Phenylpropanoid biosynthesis
[B] Lipids
Glycerolipid metabolism
Photosynthesis - antenna proteins
[B] Phytochemical compounds
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Energy Metabolism
Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Other Secondary Metabolites
Metabolism of Terpenoids and Polyketides
Metabolism of Other Amino Acid
Xenobiotics Biodegradation and Metabolism
Purine metabolism
Puromycin biosynthesis
Caffeine metabolism
beta-Alanine metabolism
Pyrimidine metabolism
Riboflavin metabolism
Folate biosynthesis
Vitamin B6 metabolism
Tyrosine metabolism
Biosynthesis of vancomycin group antibiotics
Phenylalanine, tyrosine and tryptophan biosynthesis
Phenylalanine metabolism
One carbon pool by folate
Isoquinoline alkaloid biosynthesis
Novobiocin biosynthesis
Acridone alkaloid biosynthesis
Taurine and hypotaurine metabolism
Indole alkaloid biosynthesis
Ubiquinone and other terpenoid-quinone biosynthesis
Tryptophan metabolism
Benzoxazinone biosynthesis
Histidine metabolism
Valine, leucine and isoleucine biosynthesis
Pantothenate and CoA biosynthesis
Porphyrin and chlorophyll metabolism
Clavulanic acid biosynthesis
Valine, leucine and isoleucine degradation
Cysteine and methionine metabolism
Nitrogen metabolism
D-Arginine and D-ornithine metabolism
Arginine and proline metabolism
Glucosinolate biosynthesis
Nicotinate and nicotinamide metabolism
Tropane, piperidine and pyridine alkaloid biosynthesis
D-Glutamine and D-glutamate metabolism
Biotin metabolism
beta-Lactam resistance
Penicillin and cephalosporin biosynthesis
Glutathione metabolism
Bisphenol degradation
Styrene degradation
Chlorocyclohexane and chlorobenzene degradation
Oxidative phosphorylation
DDT degradation
Ethylbenzene degradation
Tetracycline biosynthesis
Lysine biosynthesis
Lysine degradation
Propanoate metabolism
Selenocompound metabolism
Cyanoamino acid metabolism
Synthesis and degradation of ketone bodies
Glycine, serine and threonine metabolism
Pyruvate metabolism
D-Alanine metabolism
Alanine, aspartate and glutamate metabolism
Glyoxylate and dicarboxylate metabolism
Carbon fixation
Citrate cycle (TCA cycle)
Reductive carboxylate cycle (CO2 fixation)
Peptidoglycan biosynthesis
Biosynthesis of type II polyketide backbone
Biosynthesis of type II polyketide products
Phosphonate and phosphinate metabolism
Fatty acid metabolism
Fatty acid elongation in mitochondria
Butanoate metabolism
Glycerolipid metabolism
Ether lipid metabolism
Inositol phosphate metabolism
Glycolysis / Gluconeogenesis
Biosynthesis of siderophore group nonribosomal peptides
[B] Photosynthesis proteins
Sulfur metabolism
Carotenoid biosynthesis
Retinol metabolism
Flavonoid biosynthesis
Flavone and flavonol biosynthesis
Anthocyanin biosynthesis
Limonene and pinene degradation
Monoterpenoid biosynthesis
Sesquiterpenoid biosynthesis
Terpenoid backbone biosynthesis
Zeatin biosynthesis
Insect hormone biosynthesis
Steroid biosynthesis
Primary bile acid biosynthesis
Steroid hormone biosynthesis
Diterpenoid biosynthesis
Brassinosteroid biosynthesis
Secondary bile acid biosynthesis
Caprolactam degradation
Atrazine degradation
Methane metabolism
Naphthalene degradation
Aminobenzoate degradation
Dioxin degradation
Xylene degradation
Fluorobenzoate degradation
Benzoate degradation
Nitrotoluene degradation
Toluene degradation
Polycyclic aromatic hydrocarbon degradation
[B] Cytochrome P450
Metabolism of xenobiotics by cytochrom P450
Drug metabolism - cytochrom P450
Drug metabolism - other enzymes
Chloroalkane and chloroalkene degradation
Lipoic acid metabolism

5.2.2 Quantitative Analysis of Populus Organ Proteomes

In order to generate a quantitative proteome map of the different *Populus* organs, we first filtered the data to account for the stochastic nature of the peptide sampling process[168]. That is, a significant proportion of the data consists of low-abundant proteins and because accurate label-free quantitation is difficult to perform on low-abundant proteins, we applied a threshold filter for their subsequent removal. Rather than choosing an arbitrary abundance value to eliminate low abundance proteins, an empirical prevalence value was identified to obtain a "cut-off" criterion that distinguishes changes in protein expression from background noise and false positives[170]. To assess differences between organ types, only those proteins with substantive nSpC, as determined by a prevalence value (PV), were carried on to subsequent analyses. Briefly, each protein identified is given a PV, which is determined by averaging the nSpC values across all samples. Next, PVs were plotted as a histogram to graphically capture the distribution of assigned spectra, such that one could assess the cumulative spectra assigned at varying PV cut-offs. Through iterative removal of proteins below each PV cut-off, only proteins considered to be highly representative or reproducible remained. Using this approach, an ideal PV cut-off of 2.0 was determined. Applying this filter to the entire data set reduced the number of quantifiable protein groups from 7,538 to 3,242. Notably, while only ~43% of the data set remains, we retained ~98% of the total assigned nSpC values for quantitative analysis.

Using these parameters, we sought to identify the distribution of protein expression across the different organs. Hierarchical cluster analysis was applied to the 3,242 protein groups, resulting in 14 clusters that can be visualized in Figure 5.4. For hierarchical clustering, transformed data across all organ types were compared via the Fast Ward clustering algorithm using the STD option to standardize protein abundance values across all organs on a protein-by-protein basis, which essentially converts abundance values to standard deviations above or below the row mean. Proteins exhibiting similar trends across all organs were grouped into clusters and visualized by heat map to ascertain organ-specific protein representation.

Figure 5.4. Hierarchical clustering classifies protein groups by distinct localization trends. Identified protein groups above the determined prevalence value were clustered into groups based on nSpC abundance patterns across all organ types. Abundance values, ranging from -1.56 to +1.56, were calculated by converting nSpC for each protein group, averaged across all replicates, to a value representing the number of standard deviations away from the row mean. Protein groups sharing similar standardized abundance trends were then clustered into distinct families (listed top to bottom - 1, 12, 4, 8, 7, 2, 13, 10, 3, 11, 6, 14, 9 and 5) and denoted in alternate colors. Columns representing each organ-type were then clustered (bottom) based on global data set similarities.

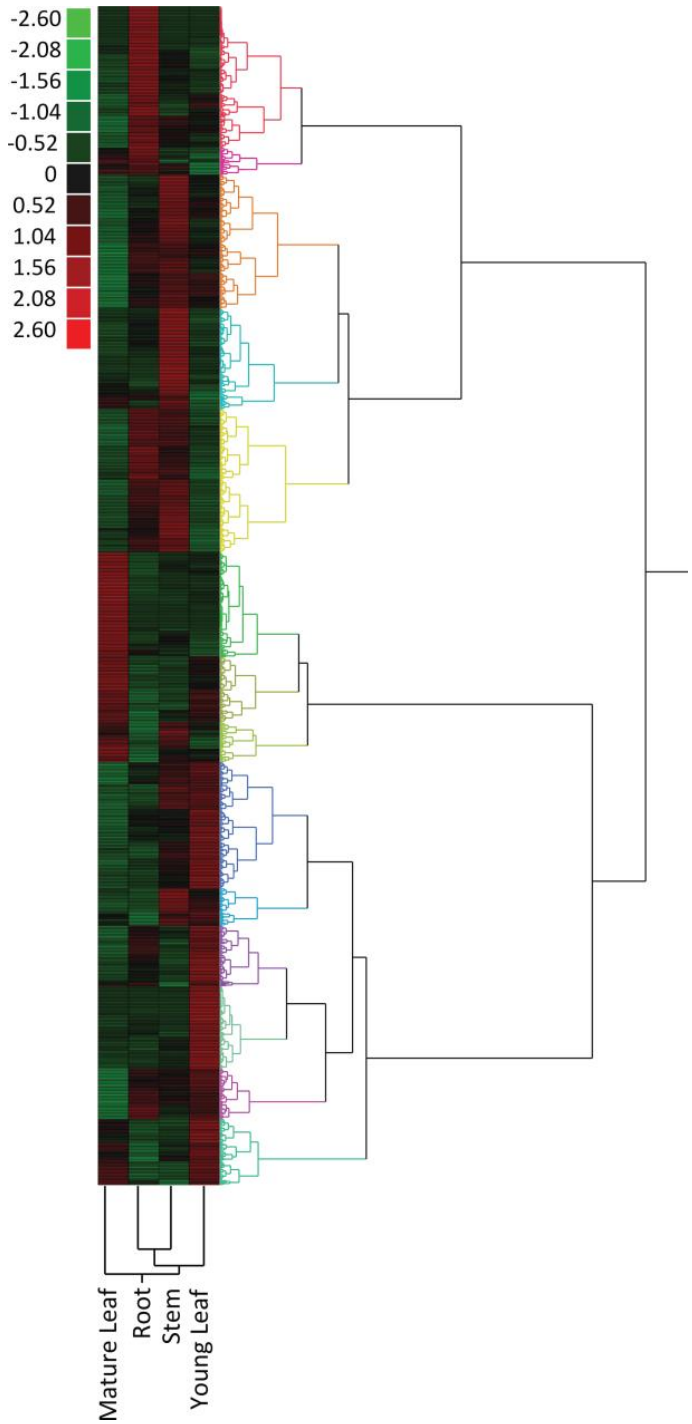
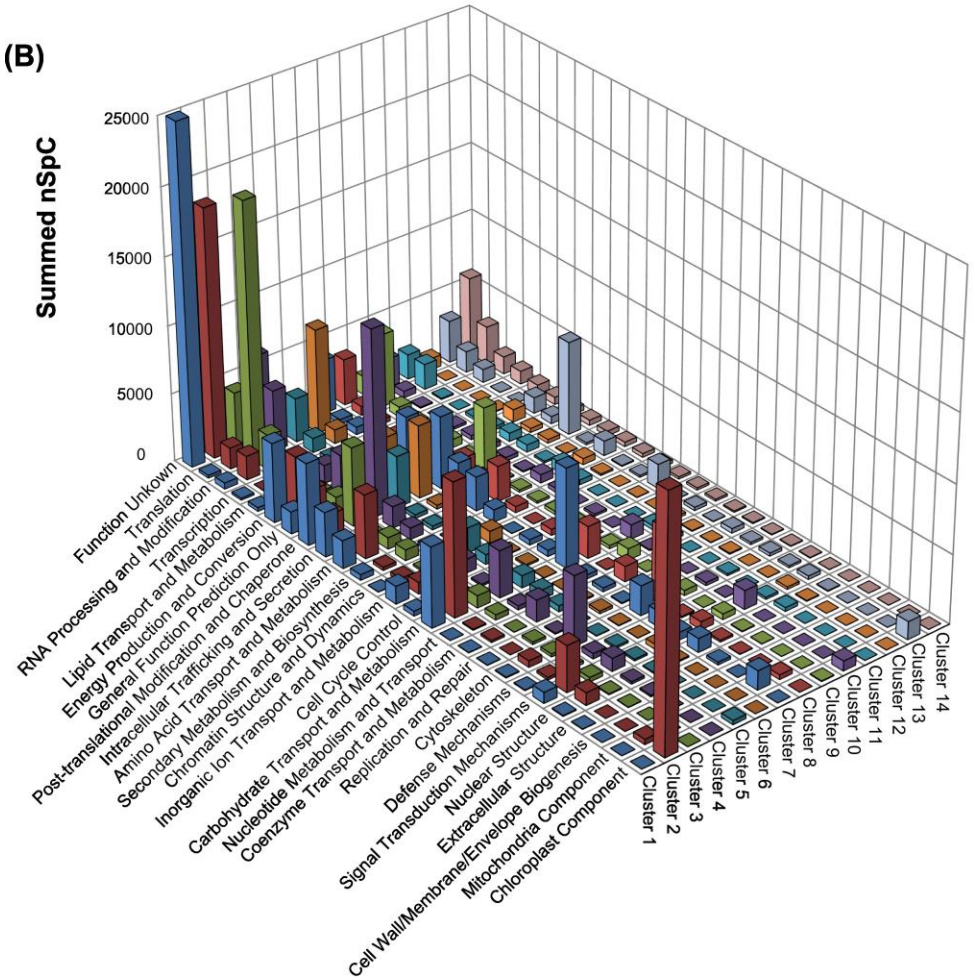

-2.60
-2.08
-1.56
-1.04
-0.52
0
0.52
1.04
1.56
2.08
2.60
Mature Leaf
Root
Stem
Young Leaf

Figure 5.5. Quantitative distribution of detected protein groups by their KOG functional classification categories for each hierarchical cluster. Protein group clusters were deconvoluted by organ-type (A) to show each organ's nSpC contribution relative to the total nSpC populating each cluster (across all organs). Table cells are color-coded based on percent contribution (green:red::low:high) in order to quickly visualize each organ's share of the total nSpC. (B) To view the functional signature of each cluster (z-axis), cluster members were classified into their respective KOG categories (x-axis), with each category's representation weighted, based on the sum of nSpC of contributing protein group members (y-axis).

(A)

Cluster	Percent Contribution of Organ nSpC			
	Mature Leaf	Root	Stem	Young Leaf
1	6.3%	66.0%	15.2%	12.5%
2	77.2%	2.4%	10.2%	10.2%
3	8.0%	16.5%	29.4%	46.1%
4	7.8%	26.2%	40.2%	25.8%
5	29.0%	8.5%	12.5%	50.1%
6	11.5%	29.3%	8.4%	43.5%
7	11.9%	37.4%	38.0%	12.7%
8	11.9%	15.3%	63.7%	9.1%
9	8.0%	32.2%	25.5%	34.2%
10	42.5%	5.9%	32.8%	18.8%
11	16.6%	11.4%	42.1%	29.9%
12	27.0%	41.9%	20.9%	10.2%
13	54.0%	4.1%	10.5%	31.4%
14	7.4%	7.7%	12.5%	72.3%

(B)



Across all clusters, the number of protein groups ranged from 395 (cluster 7) to 74 (cluster 12). While cluster membership reflects the relative diversity of protein function, the overall activity of each cluster is revealed through the relative percentages of the total assigned spectra (Figure 5.5A). With such values, the quantitative representation of each organ within a cluster can be defined.

To interpret the biological significance of each cluster, cluster membership was plotted against KOG category (Figure 5.5B). First, we examined the protein groups that were predominately expressed in only one organ: cluster 1 (root), cluster 2 (mature leaf), cluster 8 (stem), and cluster 14 (young leaf). For the set of protein groups that were predominantly expressed in roots, the three most abundant functional categories observed were "unknown function", "post-translation modification and chaperones", and "amino acid transport and metabolism". For those protein groups whose function remains unknown, an attempt to elucidate a biological role was dependent on whether a protein could be associated with a particular protein family in the Pfam database[232]. Although functional annotations based solely on family membership must be interpreted with caution, high-quality association with a protein family would, in fact, indicate what functional units are present and thus suggest a biological role. By investigating protein family membership, a functional role for the two most abundant proteins (POPTR_0013s10350.1 and POPTR_0013s10380.1) in the unknown category could be hypothesized. When searched against the Pfam database, both proteins matched with high confidence to a phosphorylase superfamily that includes 5'-methylthioadenosine phosphorylase. A previous publication[233] suggests that the ortholog represented in A. thaliana responds to changes in the level of cytokinin production in various cell types. Within plants, the phytohormone-related cytokinin is an important regulator of plant growth and, notably, 5'-methylthioadenosine, the substrate for the above-mentioned protein, has been linked to cytokinin metabolism[234]. For another set of proteins (POPTR_0008s13030.1 and POPTR_0008s13040.1), which are also among the highest expressed in the unknown category, *A. thaliana* orthologs have been shown to correlate with cytokinin levels in roots[235]. A search against the Pfam database resulted in high confident annotations for both proteins, matching against the Bet v I allergen protein

family. In fact, members of the Bet v I allergen protein family are storage proteins that occur across dicotyledonous plants[236] and have been shown to be cytokinin-binding proteins[237]. Together, the expression patterns observed, as well as annotations provided through protein family memberships, suggest biological roles impacting multiple aspects of plant development, including cell growth and sink/source relationships.

Within the set of protein groups that were predominately expressed in stem (cluster 8), the three most abundant KOG functional categories observed were "unknown function", "cytoskeleton", and "amino acid transport and metabolism". A set of the most abundant proteins shared a common biological thread; they all are involved in cell wall formation. Among this set, UDP-glucose pyrophosphorylase (UGPase; POPTR_0004s07280.1) and UDP-glucose dehydrogenase (UGDH; POPTR_0004s11760.1) were identified at similar abundance values. In plants, the enzyme UGPase is metabolically positioned at the point of sucrose synthesis/breakdown and, at this important carbon flow junction, produces UDP-glucose, which is required for sucrose synthesis or other polysaccharides, such as hemicellulose or pectin[238]. Because actively growing stem tissues (i.e., phloem and xylem) do not serve a nutritional storage role, we hypothesize that the enzyme UGDH utilizes the UDP-glucose produced by UGPase to form UDP-glucuronate, which is a precursor for hemicellulose and pectin formation, two xylem-related polymers. Lastly, the protein with the largest abundance value is an actin depolymerizing factor (ADF; POPTR_0009s03320.1) that perhaps plays a role in control of woody tissue development. In *Populus*, ADF activity is thought to be essential for the development of phloem and xylem[239].

The three most abundant KOG functional categories observed in mature leaf (cluster 2) were "unknown function", "chloroplast", and "carbohydrate transport and metabolism". A mature leaf harbors a highly active network of chloroplasts, an organelle where light energy is collected and converted into stored chemical energy that ultimately fixes atmospheric carbon dioxide to carbohydrates. Hence, fully developed leaves possess the highest photosynthetic rate, chlorophyll accumulation and respiration levels[240]. Indeed, the three most abundant proteins (ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) large subunit; Chloroplast 11241, PSAD photosystem

I reaction center subunit; POPTR_0008s15100.1 and glyceraldehyde 3-phosphate dehydrogenase; POPTR_0014s13660.1) observed in the data set reflect the main purpose of this specialized organ and substantiates its characteristic role in photosynthesis and carbohydrate metabolism.

The three most abundant KOG functional categories observed in young leaf (cluster 14) were "unknown function", "translation", and "RNA processing and modification". Unlike mature, fully expanded leaves, the more juvenile leaves appear to be ontogenetically closer to the shoot apical meristem. Rather, they appear to utilize most of their resources in active growth and development. The two most abundant proteins in young leaf were of unknown function but by investigating protein family membership with Pfam, both proteins (POPTR_0124s00210.1 and POPTR_0018s09610.1) matched with high confidence to members of the GDSL-like Lipase/Acylhydrolase family. In general, the GDSL-like lipase superfamily is thought to play an important role in the regulation of plant development and, recently, in the metabolism of cutin and wax[241-242]. In plants, cutin biosynthesis is a crucial component in the formation of outermost epidermal cell wall surface, the cuticle. Within expanding young leaves, the enzymatic mechanisms involved in the production of cutin monomers have been well-studied[243-244]. However, little progress has been made in identifying the enzymes involved in the transportation and building of the cutin matrix within the epidermal cell extracellular matrix. Lipase-type enzymes have been suggested to be involved in the cutin polymerization step within the extracellular matrix[245]. The results here, in correlation with the ANOVA analysis below, suggest that these abundant lipase proteins are involved in the formation of the plant cuticle.

5.3 Profiling Leaf Development

5.3.1 Quantitative Analysis of Populus Leaf Development

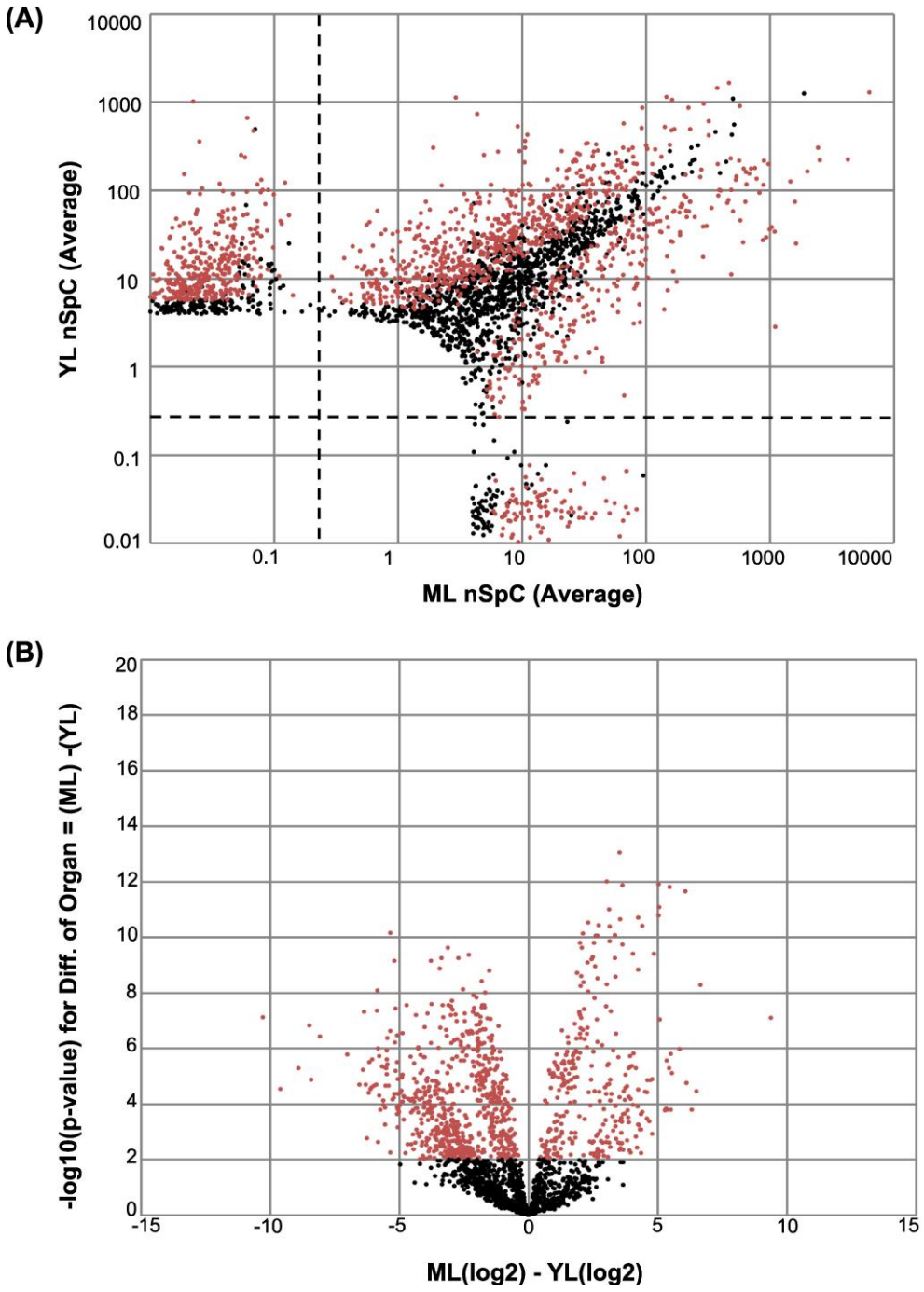
The semi-quantitative power of LC-MS/MS-based proteomics was employed to detail the proteomic differences between *Populus* leaf at two different developmental stages: young (YL) and fully expanded, mature leaf (ML). To accomplish this task, protein group normalized spectral counts (nSpC) collected across all replicates of each

leaf-type were analyzed by one-way ANOVA. Protein groups represented in the analysis include only those with significant sample-to-sample representation, as assessed by prevalence value. In total, 2881 protein groups from both young and mature leaves were statistically compared. Roughly half (1432 protein groups) were found to exhibit statistically significant ($p \leq 0.01$) differential abundance patterns, with 395 groups showing increased abundance in mature leaf compared to 1037 in young leaf (Figure 5.6A-B). These values support the proposition above that mature leaf has "settled" into its organ-specific function (photosynthesis) and thus requires a reduced complement of proteins, relative to young leaf. In contrast, young leaves are still developing, as evidenced by the up-regulation of general biosynthetic pathways (i.e., DNA synthesis, transcription, translation, etc.).

In order to better visualize the functional differences between young and mature leaves, differentially abundant protein groups from the ANOVA analysis were mapped to KEGG-derived metabolic pathways using iPath2.0. Only those *Populus* protein groups with assigned function (i.e., KEGG KO or Enzyme EC number) could be mapped, leaving out several highly abundant, differentially expressed proteins of unknown function. Despite this limitation, developmentally responsive protein groups matched to 1444 metabolic map elements (392 in mature leaf vs. 1052 in young leaf, redundant entries included) and allow for a more pathway-centric view of the functional reactions specific to leaf developmental stage (Figure 5.7A-B).

Functional enzymes exhibiting differential abundance patterns are highlighted on the pathway maps as varying degrees of either red (up-regulated in mature leaf, Figure 5.7A) or green (up-regulated in young leaf, Figure 5.7B), depending on their fold change. Protein expression that differs by a factor of 10 or more in either direction is represented by the brightest of either color.

Figure 5.7. Differential proteomic analysis of young versus mature leaf by ANOVA. Protein groups identified in young and/or mature leaf above the determined prevalence value were analyzed by ANOVA to compare the functional signature between two distinct developmental stages of leaf. (A) Protein group abundances (nSpC), averaged across all replicates (n=6) per stage, were compared between young leaf (YL, x-axis) and mature leaf (ML, y-axis) and visualized as a scatterplot. Protein groups that showed a significant ($p \leq 0.01$) difference in abundance are colored red. Dotted lines separate "effectively zero" sub-distributions from the main distribution in the top right quadrant. Proteins groups in this main distribution were identified in both developmental stages while proteins in the sub-distributions were likely found in only one stage. To further visualize the statistical metrics of the main distribution, a volcano plot (B) was constructed, comparing the LOG2 (nSpC)-based difference between both developmental stages (x-axis) to the level of statistical significance, represented as -LOG10(p-value) (y-axis). As in (A), protein groups that showed a significant ($p \leq 0.01$) difference in abundance are colored red.


(A)
YL nSpC (Average)
ML nSpC (Average)
10000
1000
100
10
1
0.1
0.01
0.1
1
10
100
1000
10000
(B)
-log10(p-value) for Diff. of Organ = (ML) -(YL)
ML(log2) - YL(log2)
20
18
16
14
12
10
8
6
4
2
0
-15
-10
-5
0
5
10
15

As a testimony to the power and accuracy of semi-quantitative proteomics, LC-MS/MS-derived protein abundance patterns highlight several contiguous pathway components, a majority of which respond in an appropriate, concerted fashion specific to leaf developmental stage. This pathway-centric view thus expands upon a general list of up- and down-regulated proteins, allowing for more complete synthesis of systems biological information. However, this is not to suggest that the latter is unnecessary, especially as only a subset of leaf stage-responsive proteins could be effectively mapped to a particular metabolic pathway.

5.3.2 Metabolic Pathway Mapping of Mature Leaf Highlights a Primary Focus on Energy Harvesting

Protein groups exhibiting increased abundance in mature leaf, relative to young leaf, substantiate the general view of a leaf as a specialized energy harvesting organ. As highlighted in box "PS" (Figure 5.8A), all major components of photosynthesis (KEGG pathway KO00195) show significant increased protein abundance in mature leaf relative to young leaf. In fact, photosynthesis is one of the most up-regulated systems present in mature leaf, an observation further evidenced by totaling the nSpC of each of the four sub-complexes: 1) Photosystem II – 5324 nSpC in mature leaf compared to 907 in young leaf (5.9x up-regulated), 2) Photosystem I – 7801 to 624 nSpC (12x up-regulated), 3) Cytochrome b6/f complex – 1873 to 370 nSpC (5.1x up-regulated) and 4) ATP synthase – 5435 to 762 nSpC (7.1x up-regulated). Furthermore, photosynthetic antenna proteins, the chlorophyll-binding components of light harvesting complexes 1 and 2 (KEGG pathway KO00196), also showed a 4.6x increase in abundance compared to young leaf (2469 vs. 534 nSpC). Taken together, mature leaf photosynthetic function was up-regulated by a factor of 7.2x relative to young leaf (22,902 vs. 3,197 nSpC).

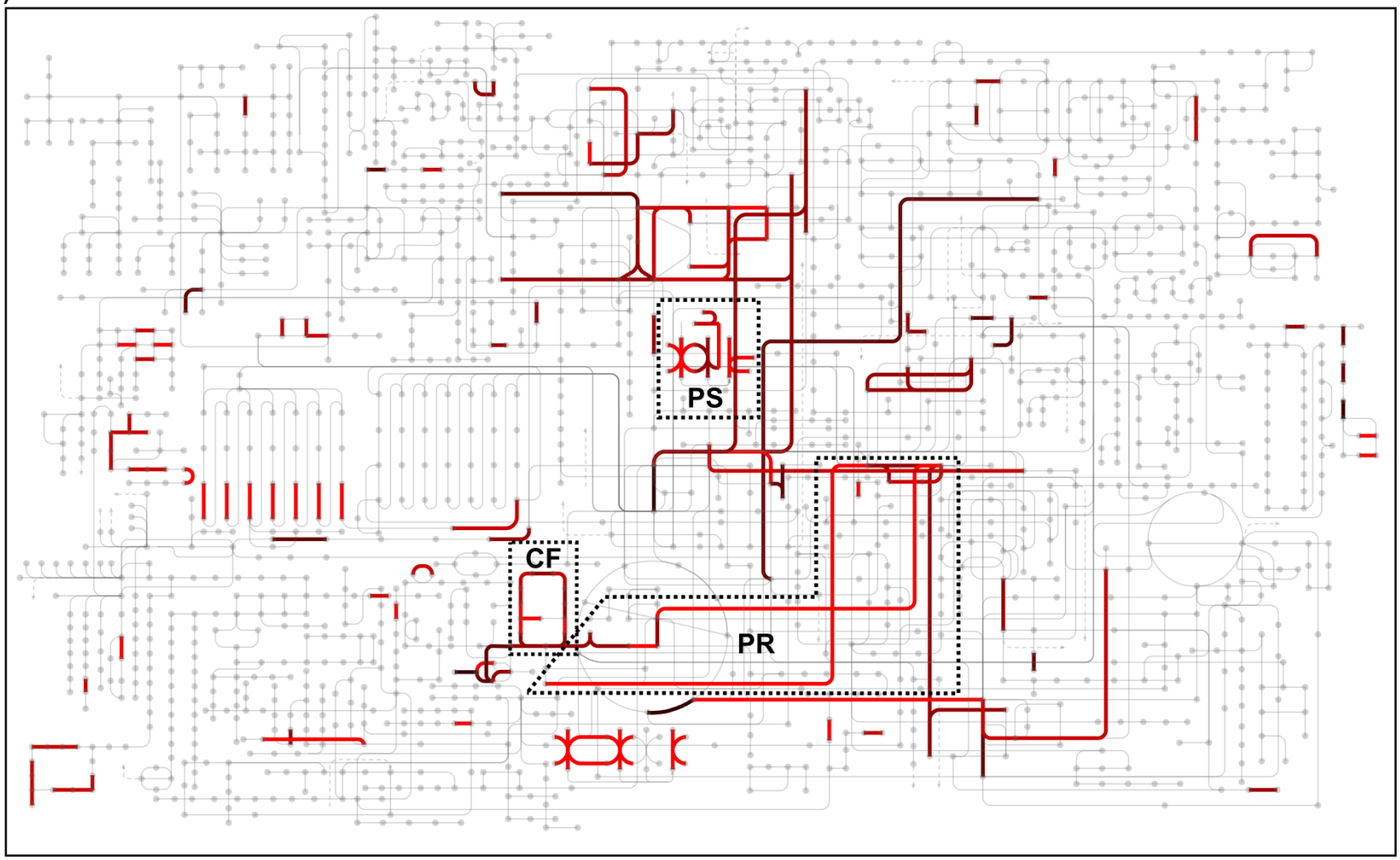
Photosynthesis is inextricably linked to carbon fixation, a process by which photonic energy harnessed from sunlight is used to replenish supplies of NADPH and ATP, both of which power the redox-based reduction of atmospheric carbon dioxide to sugar molecules. Thus, the observed increase in proteins related to photosynthesis in mature leaf must correspond to an increase in the rate of carbon fixation. As follows, enzymes relevant to the carbon fixation pathway (KO00710), which are highlighted in

box “CF” (Figure 5.8A), exhibit increased abundance in mature leaf relative to young leaf. By totaling the nSpC of the proteins involved in Calvin cycle, C3-based carbon fixation activity is up-regulated by a factor of roughly 6x (16,982 vs. 2818 nSpC in mature and young leaf, respectively), a value that is in line with the degree of photosynthesis up-regulation reported above. Furthermore, RuBisCO, the key enzyme in carbon fixation, accounted for 9,397 nSpC across both mature and young leaves, but was enriched over 4.6x in mature leaf (7,721 vs. 1,676 nSpC).

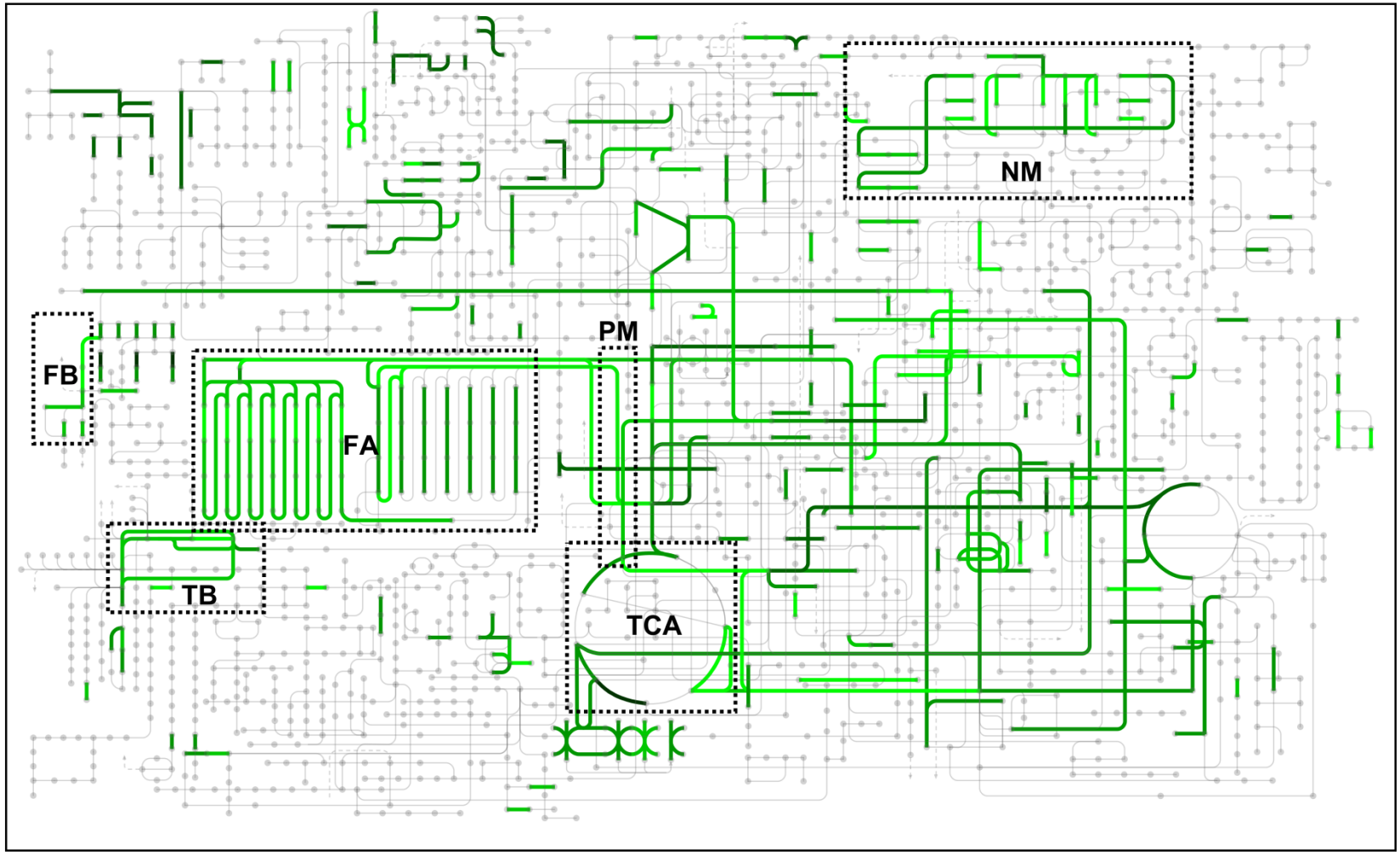
Though the primary function of RuBisCO is to fix atmospheric CO_2 to ribulose 1,5-bisphosphate (RuBP) through its carboxylase activity, this enzyme can also function as an oxygenase in a process termed photorespiration (PR). In this regard, O_2 rather than CO_2 is assimilated, leading to the production of 3-phosphoglycerate (3PG) and 2-phosphoglycolate (2PG), the latter of which must be metabolized to 3PG for re-entry into the Calvin Cycle. This complicated, multi-organelle pathway, however, equates to an expenditure of metabolic energy to both convert 2PG to 3PG and to recapture carbon (CO_2) and nitrogen (NH_4^+) lost in the process. Although the rate of photorespiration is exacerbated under hot/dry conditions, it occurs at substantial rates (~25%) even under moderate growth conditions[246]. As highlighted in box “PR” (Figure 5.8A), there was up-regulation of the photorespiration pathway (glyoxylate and dicarboxylate metabolism, KO00630), starting with RuBisCO’s oxygenase-dependent production of 2PG through its conversion to glycerate (2PG → glycolate → glyoxylate → glycine → serine → hydroxypyruvate → glycerate) and involving the necessary accessories pathways/enzymes (catalase, glutamine/glutamate cycle and tetrahydrofolate cycle) to complete the process. The only enzyme within the pathway not identified to be up-regulated was glycerate kinase. In total, PR in mature leaf was up-regulated by a factor of 5x (11,230 vs. 2251 nSpC) relative to young leaf.

Figure 5.8. Up-regulated metabolic pathways as dictated by *Populus* leaf developmental stage. Proteins exhibiting differential abundance patterns (ANOVA; $p \leq 0.01$) across both young and mature leaf were mapped to KEGG pathways using iPath v.2.0 and color-coded to indicate the degree of protein abundance differences between each developmental stage. (A) Proteins with significantly increased abundance in mature leaf are labeled in red with brighter shades indicative of larger differences. Highlighted pathways (dashed boxes) include photosynthesis (PS), carbon fixation (CF), and photorespiration (PR). (B) Proteins with significantly increased abundance in young leaf are labeled in green with brighter shades indicative of larger differences. Highlighted pathways include nucleotide metabolism (NM), flavonoid biosynthesis (FB), fatty acid metabolism (FA), pyruvate metabolism (PM), terpenoid biosynthesis (TB), and TCA cycle (TCA).

(A)



(B)



These three major mature leaf-enriched metabolic pathways constitute a proof-of-concept with regard to the LC-MS/MS platform described in this paper. As mentioned above, and further corroborated by these proteomic data, mature leaf appears to have "settled" into its primary function. Other less complete pathways were also found to be up-regulated, with a portion of them seemingly involved in reacting to oxygenic stress, most likely induced by the photosynthetic process itself. For example, L-ascorbate peroxidase (EC:1.11.1.11, KO00434) was up-regulated in mature leaf by 3.9x (223 vs. 57 nSpC) while a 2-cysteine peroxiredoxin (EC: 1.11.1.15, KO03386) was up-regulated by a factor of 2.6x (631 vs. 240 nSpC). Both enzymes are known to detoxify reactive oxygen species, with the latter previously shown to be targeted to chloroplasts to provide prevent photooxidative damage to the photosynthetic membrane[247-248]. Furthermore, three enzymes in the pathway for carotenoid and xanthophyll biosynthesis were also slightly up-regulated in mature leaf (lycopene beta-cyclase [KO06443], zeaxanthin epoxidase [KO09838] & 9-cis-epoxycarotenoid dioxygenase [KO9840]) with modest nSpC differences. In fact, the spectral counts from mature to young leaf totaled 26 vs. 0, supporting the premise that the mature leaf has enhanced photosynthetic capabilities, including pigments that modulate harvested photonic energy in periods of high light and act as antioxidants, playing lipid-protective roles in periods of high photosynthetic/oxygen generating activity[249-250].

5.3.3 Metabolic Pathway Mapping of Young Leaf Highlights a Primary Focus on Growth and Development

Young leaf engages in most, if not all, of the photosynthetic-dependent pathways detailed above, albeit with reduced abundance relative to mature leaf. The reduced level of photosynthetic-related proteins is countered by a systematic increase in protein abundance in several major general biosynthetic pathways, consistent with the fact that a young leaf is primarily actively growing and secondarily photosynthesizing. This perhaps intuitive observation is apparent upon viewing the pathway map in Figure 5.8B. Relative to mature leaf, young leaf shows increases in several metabolic pathways including terpenoid biosynthesis ("TB" | KO00900 | ~2.3x | 177 vs. 78 nSpC), flavonoid biosynthesis ("FB" | KO00941 | ~1.5x | 58 vs. 39 nSpC), pyruvate metabolism ("PM" |

KO00620 | ~1.6x | 1443 vs. 888 nSpC), TCA cycle flux ("TCA" | KO00020 | ~2.0x | 1282 vs. 630 nSpC), fatty-acid metabolism ("FA" | KO00061, KO00062, KO00071 | ~3.3x | 321 vs. 98 nSpC) and nucleotide metabolism ("NM" | KO00230, KO00240 | ~3.3x | 735 vs. 221 nSpC).

Although each of these pathways, comprised of their respective pools of differentially expressed proteins, were found to be up-regulated in young leaf, their spectral representation is less abundant relative to observation for regulatory pathways, including transcription (KO03020, KO03022, KO03040), translation (KO03010, KO00970, KO03013, KO03015, KO03008) and protein folding/sorting/degradation (KO03050, KO03018, KO03060, KO04120, KO04141, KO04130, KO04122). Proteins involved in the Populus translational apparatus were by far the most abundant in young leaf, exhibiting a ~3.6x increase in abundance (15,691 vs. 4,342 nSpC) relative to mature leaf. This increase was corroborated by significant increases in both transcription (~3.0x, 1,875 vs. 630 nSpC) and protein folding/processing (~3.6x, 6,327 vs. 1,731 nSpC). Taken together, and including a modest increase observed in DNA metabolism (96 vs. 4 nSpC), young leaf regulatory pathways were up-regulated by a factor of roughly 3.6x (23,989 vs. 6,707 nSpC). The measured increase in protein abundance observed for members of these specific regulatory pathways solidify the general observation that young leaf's primary function is growth that is fueled by moderate levels of photosynthesis and carbon fixation.

5.4 Conclusions

As demonstrated, the optimization of sample preparation and instrumentation provided the necessary depth to generate detailed proteome maps of each principle organ. By combining the proteome maps, we generated a proteome atlas that offers a broad overview of the *Populus* proteome and also the ability to zoom in on specific biological features, such as biological pathways and individual proteins. At the highest level, the 'zoomable' proteome atlas facilitated the identification of biological networks that were unique to a particular organ but also allowed us to identify the network of proteins that define the *Populus* core proteome. Looking at a pathway-level perspective, we identified

several regions of biological pathways were that were unique to a specific organ. The proteins that create these discrete regions are likely those positioned at key metabolic junctions or branching points, where an organ-specific protein is responsible for altering the flow of biological information. For example, we identified a set of proteins unique to root, which mapped to the sucrose synthesis/breakdown pathway, and a set of proteins unique to stem, which also mapped to the sucrose synthesis/breakdown pathway. In this example, the proteins unique to root facilitate the transformation of sucrose into storage reserves (i.e., starch), whereas the proteins unique to stem are metabolically positioned to transform sucrose into cell wall components. Hence, these proteins are capable of directing carbon flow for either nutritional storage or cell wall construction. Clearly, the pathway-level perspective will become an indispensable viewpoint when trying to make reliable predictions about the response of the cellular system to environmental perturbations and experimental manipulations.

At the protein-level, in addition to identifying proteins with a known biological function, a large percentage of each organ proteome consisted of proteins with no known function. Though specific biological roles were not determined for these proteins in this present study, general observations (i.e., organ location, differential regulation, etc.) outlined here provide hypotheses for further interrogation. In addition to providing qualitative data describing the protein complement of each organ, the collected data also contains semi-quantitative information, reflecting the underlying functional processes and mechanisms in each organ-type as weighted by a conservative estimate of protein abundance. To demonstrate the precision and comprehensiveness of this semi-quantitative approach, we explored proteomic differences between the same organ-type during two growth stages, young and mature leaf. As detailed above, mature leaf appears to function primarily in an autotrophic role consisting of energy generation via its photosynthetic apparatus and reduction of CO_2 to sugar via its carbon fixation pathway. Though other photosynthesis-related pathways were up-regulated in mature leaf (i.e., photorespiration and defense against photosynthetically-derived reactive-oxygen species), proteins involved in photosynthesis and carbon fixation constitute the majority of the quantitative signal. On the other hand, biosynthetic and regulatory functions were

relatively up-regulated in young leaf. Even though proteins/pathways for photosynthesis, carbon fixation and photorespiration were detected in young leaf, they were less represented relative to mature leaf. This information suggests that young leaves partition resources between growth and energy production. These observations and data provide a "proof-of-concept" with regard to our 2D-LC-MS/MS platform and suggest the biological validity of these pathway-centric comparisons, opening the door for future hypothesis-driven inquiries into Populus and other complex organisms. Obviously caution must be exercised when interpreting these semi-quantitative results, as only one biological replicate was available for statistical assessment. Clearly the inclusion of more biological replicates (3-5) would improve the statistical framework of this discovery-based approach. Nevertheless, the workflow discussed here provides an intellectual springboard for future targeted, quantitative approaches.

CHAPTER 6

MOVING AWAY FROM THE REFERENCE GENOME: EVALUATING SINGLE AMINO ACID POLYMORPHISM IDENTIFICATIONS FROM A PEPTIDE SEQUENCING TAGGING APPROACH FOR THE GENUS *POPULUS*

All of the data presented below has been adapted from the following submitted journal article:

Paul Abraham, Rachel Adams, Gerald Tuskan, Robert Hettich. "Moving Away from the Reference Genome: Evaluating Single Amino Acid Polymorphism Identifications from a Peptide Sequencing Tagging Approach for the Genus *Populus*". *Journal of Proteome Research (**In review**)*. Sample preparation and mass spectrometry experiments were performed by Paul Abraham. RNA-Sequencing derived proteome databases were generated by Xiaojing Wang at the University of Vanderbilt. Biological data analysis was performed by Paul Abraham.

6.1 Introduction to *Populus* Genetic Diversity

In contrast to other plant models such as *Arabidopsis* and rice, which are predominately self-fertilizing and consequently maintain low levels of allelic polymorphism, the *Populus* genus is primarily composed of dioecious, self-incompatible woody plants[251]. Obligate outcrossing combined with wind-pollination and prolonged reproductive life generates highly heterozygous populations with low levels of linkage disequilibrium. This mating system results in high levels of gene flow and extensive nucleotide variability within and across *Populus* species, providing an excellent model system to investigate the relationship between naturally occurring single-nucleotide polymorphisms (SNPs) and phenotypic variation[117].

Through association genetics, the discovery of nucleotide variations among genotypes has the potential to reveal allelic polymorphisms underlying complex, adaptive traits. SNPs can be located either within a protein-coding region or outside coding regions. On average, SNP frequency in protein-coding regions is high in forest trees,

generally in the order of 1 per 1000 base pairs, SNP frequency in *Populus* is somewhat higher, with an estimate of 1 SNP every 200 base pairs[114]. Nucleotide polymorphisms that occur inside coding regions may (non-synonymous) or may not (synonymous) change the amino acid sequence of the corresponding protein. Because synonymous changes are largely invisible to selective pressure and have few biological implications, they are categorized as silent nucleotide variations. On the other hand, non-synonymous changes can be under strong selective pressure and because they can directly impact gene function, they are the primary focus of most physiological or pathological association studies[252].

Although SNPs in *Populus* have been extensively studied over the past decade, little attention has been paid to single amino acid polymorphisms (SAAPs) of proteins at the proteome level. In fact, only a few efforts have been made to survey SAAPs across the *Populus* proteome. As highlighted in Chapter 3, tandem mass spectrometry (MS/MS)-based proteomics was employed for the large-scale analysis of SAAPs. In general, the available protein databases used for such studies are incomplete with respect to sequence variation information. Without taking SNP variations into account, proteomic investigations generally fail to identify any protein form containing a SAAP. Therefore, we appended predicted protein sequence variations to the original database in order to detect novel protein forms. The main disadvantage of this approach, however, is that *a priori* knowledge of SNPs is required. Moreover, this approach is preconditioned on both the coverage and quality of the predictions when they are available. Therefore, we argue that a more attractive approach considers unexpected single amino acid polymorphisms.

The high-throughput discovery of protein sequence variants (truncations, post-translational modifications, or mutations), especially unexpected variants, has seen tremendous advancements in recent years[253]. Many database searching algorithms have been recently designed to effectively identify unanticipated (blind) sequence variants at a global level. One class of such algorithms uses *de novo* sequencing in order to infer full-length peptide sequences from tandem mass spectra without requiring a sequence reference database[52, 254-255]. A strength of this approach is that the concept of variant peptides is not relevant; each spectrum is given an equal opportunity to match any

combination of amino acids, regardless of whether the researcher anticipated detecting the sequence or not. This technique, however, greatly increases the number of candidate peptides compared to each spectrum, consequently incurring not only significant costs to processing time but also unacceptable false discovery rates (FDR)[256]. In addition, mass spectrometrists have developed and routinely used a hybrid approach between traditional database searching and *de novo* approaches: here peptide sequence tagging (PST) algorithms can detect unexpected sequence variants as extensions of partial sequences identified from a database[54-55, 257-258]. In particular, the proteome informatics group led by David Tabb recently released a two-step methodology involving the DirecTag algorithm[57] for highly accurate PST tag generation, followed by the TagRecon software[162] for the detection of peptide sequence variants through tag reconciliation. In brief, short sequence "tags" are directly inferred from a tandem mass spectrum and then tags are automatically reconciled against representative peptides from a protein database while making allowances for unexpected mass shifts (i.e., mutations and post-translational modifications). PSTs serve as a filter to effectively reduce the number peptide-spectrum matches being scored, which in turn improves costs in processing time, sensitivity, and specificity[41].

To evaluate a peptide sequence tagging approach for *Populus* with the ultimate goal of globally identifying unknown SAAPs, we employed DirecTag and TagRecon software. Using the state-of-the-art LTQ-Orbitrap-Pro platform, we profiled and compared two genotypes of *P. trichocarpa* and revealed a large number of unexpected SAAPs that would have otherwise been missed by a traditional database search. The sequence variants leveraged from TagRecon demonstrates the value of using peptide sequence tagging algorithms to interrogate proteomics data sets, provided that a SAAP location could be confidently identified. Therefore, while our initial aim was to comprehensively identify SAAPs, we focused on our most abundant sequence variant to show that confident site localization remains an important yet challenging task. Since others have shown that HCD fragmentation improves the coverage of peptide sequences overall, in particular for tryptic peptides up to 15 amino acids in length, we exploited HCD fragmentation to further refine a subset of the dataset.

6.2 Peptide Identification Using a Standard Database Algorithm

For this study, proteome extracts from three tissues (leaf, root, and stem) were harvested from two *P. trichocarpa* genotypes, 'DENA' and 'VNDL', and analyzed in triplicate on an LTQ-Orbitrap-Pro mass spectrometer. Using standard parameters, the collected tandem mass spectra (MS/MS) were searched with MyriMatch[63] against the *P. trichocarpa* v3.0 protein database and supplemented with the chloroplast and mitochondrial proteomes). We employed IDPicker[164] to filter the resulting peptide-spectra matches at a maximum FDR of 2% (PSM level) and assemble peptides into a list of proteins (Table 6.1). Overall, 69,613 distinct peptide sequences were detected across the entire MyriMatch dataset. Because a considerable portion of the observed peptides are shared among multiple proteins, assigning peptides to their respective proteins is a considerable challenge in *Populus*. As highlighted in previous chapters, we recommend addressing this by incorporating additional supporting information (i.e., sequence homology) to better infer the existence of proteins in the sample. Therefore, proteins sharing 90% or more sequence identity within the *Populus* database were collapsed into protein groups.

Of the original 25,550 redundant proteins observed, a total of 9,601 protein groups were identified and of those, 3,399 were singletons (i.e., one-membered groups). Because both genotypes were grown under identical growth conditions, we expected to observe substantial overlap in the proteins that were expressed in both genotypes. Indeed, the measured proteins for both 'VNDL' and 'DENA' shared a high level of overlap (~80%). For the purpose of evaluating the depth of coverage achieved, we compared the number of protein groups identified against the results in Chapter 4, which at that time had achieved the deepest proteome coverage in the genus *Populus*. Overall, >2,000 additional protein groups were detected in the current study.

6.3 Identification of Sequence Variants Using Peptide Sequencing Tagging

6.3.1 Experimental Workflow and Results

Sequence variations, manifested by single amino acid polymorphisms, provide clues to the genetic structures that induce a pathological or physiological trait. To our knowledge, SNPs are widely measured at the transcriptome level, but rarely at the proteome level. For the reasons outlined above, we employed a peptide-sequence tagging approach to identify SAAPs in *Populus*.

Figure 6.1 illustrates the three-step experimental workflow used to identify unexpected sequence variants in *Populus*. The first step uses the MyriMatch search engine to identify a confident list of proteins (no unexpected sequence variants considered) for each biological sample (Figure 6.1A), and IDPicker was employed to ensure only confident identifications were retained. This step serves to dramatically reduce the candidate list of proteins (a subset FASTA database) for the blind search that follows, with the purpose of improving processing time, sensitivity and specificity of the analysis. In the second step, DirecTag infers sequence tags from the MS/MS scans from each raw file, followed by TagRecon mass matching the inferred sequences to the subset protein database while making allowances for unanticipated mass shifts in peptides. IDPicker was employed to filter the resulting peptide-spectra matches at a maximum FDR of 2% (PSM-level) and assemble peptides into a list of proteins (Table 6.2). For the final step, peptide-spectrum matches observed in MyriMatch and TagRecon were compared to obtain a final data set of the highest quality (Figure 6.1B). In addition, several proven attestation principles[165] were applied to further validate peptide sequence variants (See Chapter 2.2.2). Table 6.3 presents a summary of the attested results after merging the data from the two database search engines.

In general, the percentage of identified peptides (frequency) and spectra (abundance) containing a sequence variant did not seem dependent on the genotype (Table 6.4).

Table 6.1. Summary of MyriMatch results.

	VNDL Leaf Replicate 1	VNDL Leaf Replicate 2	VNDL Leaf Replicate 3	DENA Leaf Replicate 1	DENA Leaf Replicate 2	DENA Leaf Replicate 3
spectra	129980	126378	127260	138884	132424	135239
distinct peptide matches	28443	22275	27063	25348	32803	30095
distinct peptides	22497	17868	21323	19790	25307	23524
proteins	14099	11557	13061	11708	14200	13539
protein groups	7420	6044	6935	6168	7559	7187
modifications	10271	10669	10110	10599	11061	10826
protein FDR %	0.33	0.24	0.25	0.63	0.58	0.37

	VNDL Stem Replicate 1	VNDL Stem Replicate 2	VNDL Stem Replicate 3	DENA Stem Replicate 1	DENA Stem Replicate 2	DENA Stem Replicate 3
spectra	142069	146643	137313	163798	159356	160969
distinct peptide matches	31154	31381	29840	33754	30900	37531
distinct peptides	25107	25215	24361	26422	24354	29064
proteins	15172	15307	15241	15016	14330	16154
protein groups	8036	8007	8049	8064	7603	8730
modifications	10075	10583	14501	11551	10512	11472
protein FDR %	0.28	0.27	0.17	0.63	0.40	0.51

	VNDL Root Replicate 1	VNDL Root Replicate 2	VNDL Root Replicate 3	DENA Root Replicate 1	DENA Root Replicate 2	DENA Root Replicate 3
spectra	169237	157369	166832	179736	164198	161507
distinct peptide matches	34058	35420	37109	40740	42295	38893
distinct peptides	27063	27862	28890	31305	32679	29085
proteins	14618	14880	15228	16062	16607	14791
protein groups	7977	8159	8369	8880	9218	8180
modifications	17662	15835	17927	16377	15055	14990
protein FDR %	0.51	0.44	0.59	0.68	0.45	0.41

Table 6.2. Summary of TagRecon results.

	VNDL Leaf Replicate 1	VNDL Leaf Replicate 2	VNDL Leaf Replicate 3	DENA Leaf Replicate 1	DENA Leaf Replicate 2	DENA Leaf Replicate 3
spectra	143205	136774	138469	150600	144022	145535
distinct peptide matches	34976	27283	32969	31543	39729	36090
distinct peptides	23559	18704	22228	20879	25970	24109
proteins	13290	10919	12474	11168	13418	12792
protein groups	7283	5922	6814	6099	7349	6993
modifications	44432	42000	42629	43565	44674	42356
protein FDR %	1.13	1.12	1.20	1.43	1.58	1.25

	VNDL Stem Replicate 1	VNDL Stem Replicate 2	VNDL Stem Replicate 3	DENA Stem Replicate 1	DENA Stem Replicate 2	DENA Stem Replicate 3
spectra	131679	152409	143157	170676	164692	167969
distinct peptide matches	32533	35810	33765	39337	35981	43259
distinct peptides	23210	25433	24787	27120	24878	29521
proteins	13749	14313	14253	14177	13505	15343
protein groups	7349	7739	7753	7902	4738	8507
modifications	37141	43671	42201	47022	43762	46903
protein FDR %	1.08	1.22	1.16	1.75	1.47	1.59

	VNDL Root Replicate 1	VNDL Root Replicate 2	VNDL Root Replicate 3	DENA Root Replicate 1	DENA Root Replicate 2	DENA Root Replicate 3
spectra	177849	164947	173479	185879	169761	173503
distinct peptide matches	39845	40813	42266	46054	47636	45437
distinct peptides	28139	28701	29500	31773	33120	30106
proteins	13903	14123	14393	15133	15780	14125
protein groups	7854	7975	8177	8636	9000	8071
modifications	52902	48578	50750	53601	48189	52602
protein FDR %	1.93	1.66	1.89	1.93	1.83	1.81

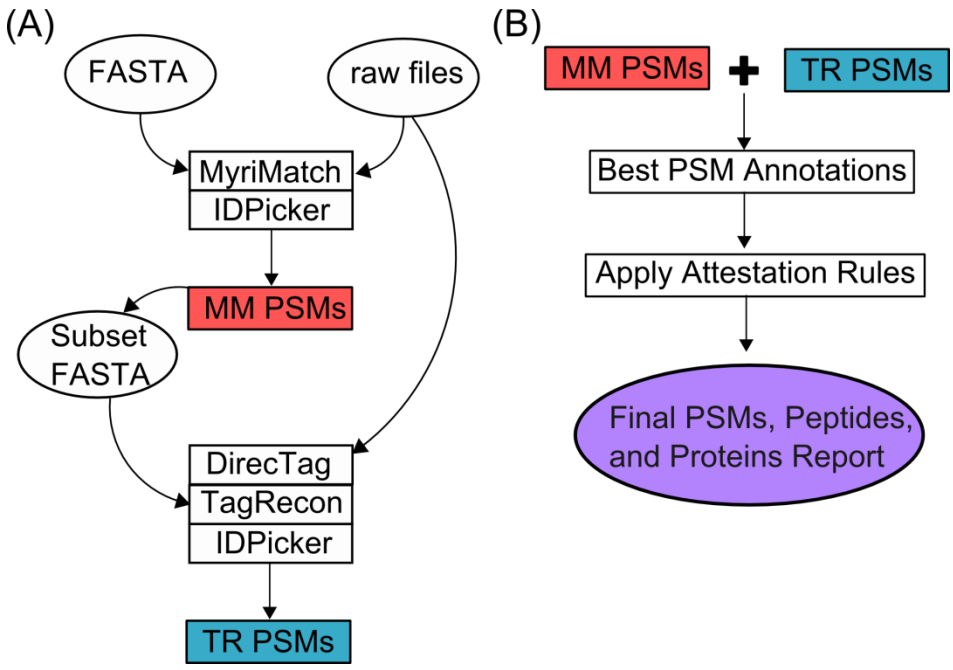


Figure 6.1. Computational workflow for the identification of peptide sequence variants. This flowchart illustrates the three-step strategy used to identify peptide sequence variants from a DirecTag and TagRecon approach. (A) First, a traditional database search was performed by MyriMatch to confidently identify a list of proteins, which was used to minimize the search database. In the second step, the DirecTag software provides an inferred sequence tag for every MS/MS spectra. TagRecon then reconciles the inferred sequence tags against the subset database to identify unexpected mass shifts in peptides sequences. Following every search, the IDPicker software applied a variety of score combinations to filter the resulting identifications at 5% FDR (B) The MyriMatch and TagRecon results were compared to identify the best PSM for every MS/MS spectra. The data set was further refined by validating every sequence variant with proven attestation rules.

Table 6.3. Results after merging MyriMatch and TagRecon data sets.

	VNDL Leaf	DENA Leaf
averaged spectra	144117	153420
summed spectra	432352	460259
peptides	23495	25200
distinct peptides	20944	22694
proteins	13688	13530
protein groups (≥90%)	4662	4643
peptide variants	2362	2800
distinct peptide variants	1834	2161
peptide variant summed spectra	31045	31366

	VNDL Stem	DENA Stem
averaged spectra	152954	176356
summed spectra	458863	529069
peptides	26870	27976
distinct peptides	24610	25705
proteins	15725	15480
protein groups (≥90%)	5500	5374
peptide variants	2142	2291
distinct peptide variants	1744	1822
peptide variant summed spectra	24512	24784

	VNDL Root	DENA Root
averaged spectra	180226	182388
summed spectra	540677	547164
peptides	29426	33766
distinct peptides	27380	31099
proteins	15210	16476
protein groups (≥90%)	5334	5827
peptide variants	1934	2257
distinct peptide variants	1537	1790
peptide variant summed spectra	20047	22205

Table 6.4. Frequency and abundance of sequence variants in *Populus*.

genotype/organ	peptides	spectra	frequency	abundance
VNDL/Leaf	2362	31045	11.3%	7.2%
VNDL/Stem	2142	24512	8.7%	5.3%
VNDL/Root	1934	20047	7.1%	3.7%
DENA/Leaf	2800	31366	12.3%	6.4%
DENA/Stem	2291	24784	8.9%	4.7%
DENA/Root	2257	22205	7.3%	4.1%

We identified a total of 6,653 peptide sequence variants (12% of total identified peptides); 4,391 and 4,900 sequence variants for 'VNDL' and 'DENA', respectively. Overall, these sequences mapped to 22,067 proteins and 8,088 protein groups, which means a peptide sequence variant was identified in 86% of the proteins observed and 84% of protein groups. Although the percentage of peptide sequence variants identified seems relatively small, this can be explained by the experimental limitations of the approach. In general, the median sequence coverage observed in shotgun proteomic experiments employing a trypsin-based schema is often between 20-25%[73]. Consequently, we anticipated a limited sampling of SAAPs across individual proteins. Nevertheless, we identified a sequence variant for nearly every protein detected. Future studies may be warranted to specifically focus on achieving maximal sequence coverage by modifying the experimental strategy to incorporate multiple proteases[259-260], which would provide more specificity to the frequencies of SAAPs per protein.

6.3.2 Types of Variants

The procedure described above identified a total of 76 types of sequence variants (each type denoting an amino acid with a mass shift corresponding to a mutation). Noticeably, the occurrence of variants in both genotypes is similar (Pearson correlation = 0.99). Of those listed, the top 20 most abundant have been highlighted in Table 6.5.

Peptides and fragment ions containing an oxidation mass shift (+15.99 Da) were the most prevalent variant type, representing ~38% of the total assigned spectra for variant peptides. While this observation may suggest the two most prominent SAAPs are Ala→Ser and Phe→Tyr, we critically evaluated the results by validating each variant through manual verification of the MS/MS spectra. In the course of this inspection, we observed that the site of +16 Da mass shifts were often in close proximity to a methionine residue (see Figure 6.2), which is frequently oxidized during sample processing. Correspondingly, the site of a ΔA=32 Da mass shift, which can correspond to double-oxidation event or two singly-oxidized alanine residues, was also often found near methionine residues. Therefore, the source of the most frequent and abundant SAAPs could perhaps be explained away as a "shadow" of the most common sampling processing artifact.

Table 6.5. Top 20 sequence variants observed in *Populus*.

Mutations Identified	UniMod Annotation	VNDL Peptides	DENA Peptides	VNDL Summed Spectra	DENA Summed Spectra
A[16]	Ala->Ser	978	1147	22439	22459
F[16]	Oxidation; Phe->Tyr	345	387	7024	7424
M[-3]	Met->Lys; Met->Gln	247	261	5399	5351
N[-27]	Asn->Ser	201	229	3366	3437
V[14]	Val->Xle	153	169	2434	2998
S[14]	Methyl; Ser->Thr	44	67	2353	798
A[32]	Ala->Cys	106	64	2206	2091
V[2]	Val->Thr	130	121	1872	1890
A[28]	Ala->Val	69	155	1811	1355
S[42]	Acetyl; Ser->Glu	99	89	1424	1417
S[41]	Ser->Lys; Ser->Gln	82	81	1268	1344
G[30]	Gly->Ser	59	71	1217	1337
V[32]	Val->Met	74	106	1192	1227
S[-16]	Deoxy; Ser->Ala	82	82	991	1480
N[15]	Methyl+Deamidated; Asn->Glu	70	49	916	1129
L[-14]	Xle->Val	41	69	861	567
N[-13]	Asn->Thr	27	76	861	1192
G[14]	Gly->Ala	73	60	856	843
T[-30]	Thr->Ala	57	65	766	832
V[-28]	Val->Ala	68	27	741	1289

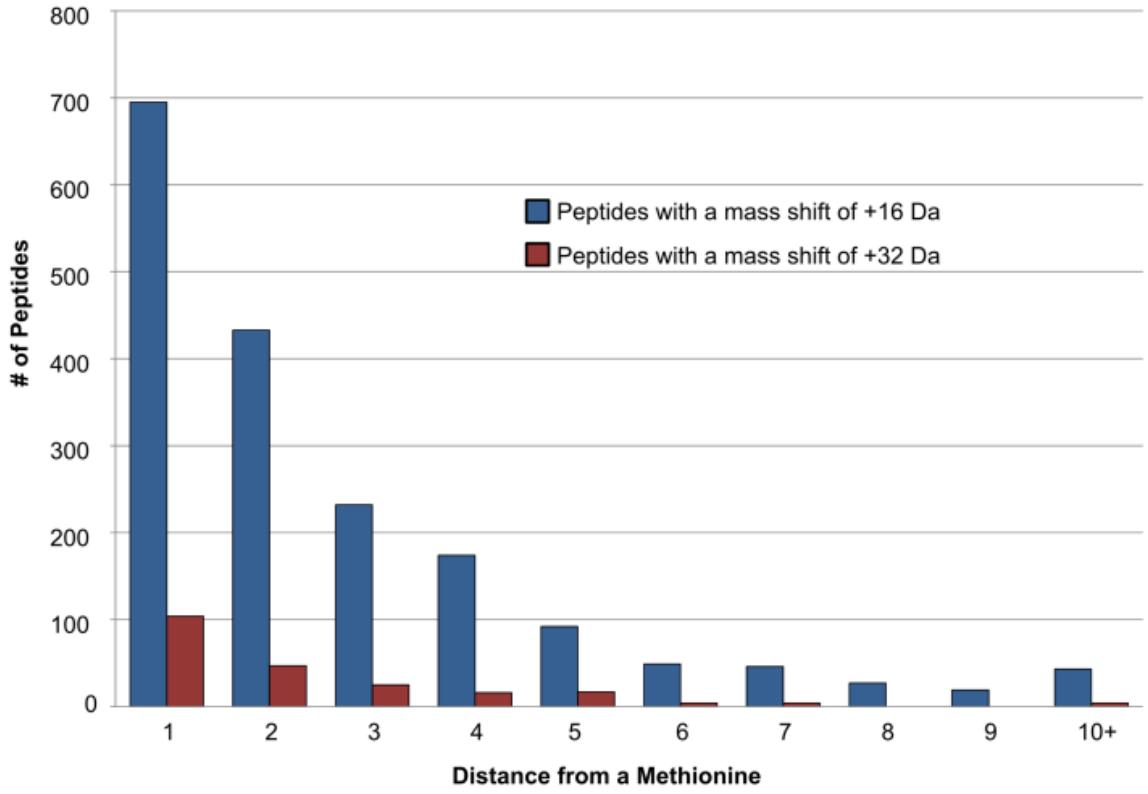


Figure 6.2. Proximity of +16 Da and +32 Da mass shifts to methionine residues. Detected peptides containing a methionine residue and either a +16 Da (blue) or +32 Da (red) mass shift on a non-methionine residue were plotted. This frequency distribution illustrates the degree of adjacency between the modification site and neighboring methionine residues.

Though the presence of a mass shift changes the ion fragmentation pattern of the corresponding ions, the fragmentation process is often incomplete. Some mass shifts will lead to unique fragmentation patterns, enabling a site to be unambiguously located. On the other hand, a mass shift that can occur at adjacent residue sites can introduce ambiguity and lead to incorrect localization; the candidate peptide variants will have similar theoretical fragmentation patterns and thus similar statistical scores. As the distance between the two sites increases, complementary site-determining b- and y-type ions together should increase a scoring algorithm's ability to mitigate the ambiguity (Figure 6.3). Therefore, we objectively evaluated how this ambiguity diminishes as the adjacency decreases.

The analysis was constrained to 'DENA' leaf samples, which contained the highest frequency and abundance of ΔA=16 Da mass shifts. Since high mass accuracy of fragment ions can help unambiguously annotate fragment ion peaks, a MS run using a 'high-high' strategy, which means full scans (MS) and tandem mass spectra (MS/MS) are detected in the Orbitrap analyzer at high resolution and high mass accuracy, was simultaneously evaluated with the a MS run that acquired MS/MS scans in the ion trap ('high-low'). The collected spectra were searched by MyriMatch using a directed method (See Materials and Methods); only a user-defined mass shift was considered. For both MS runs, two directed searches were performed: either a methionine (+16 Da) or an alanine (+16 Da) was allowed as a dynamic modification. By searching for the modifications independently, the search algorithm interpreted each spectrum, identified the mismatch region containing a permissible modification and determined the most probable position of the mass shift on either the methionine or alanine. This approach enabled the identification of spectra that were annotated similarly, having the same underlying peptide sequence but differing by the location of the mass shift, either on a methionine or a neighboring alanine. For discussion purposes, these spectra will be referred as 'contentious spectra' (CS). In total, the MS searches identified nearly the same number of CS for each analysis strategy – 37,776 and 38,399 for high-high and high-low, respectively.

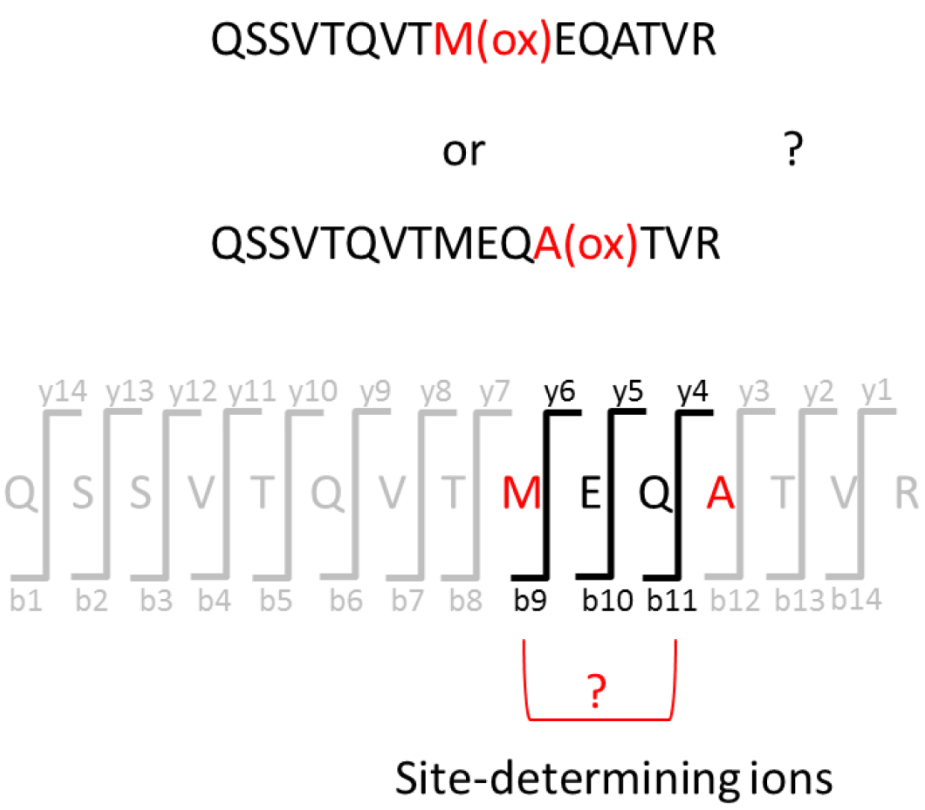
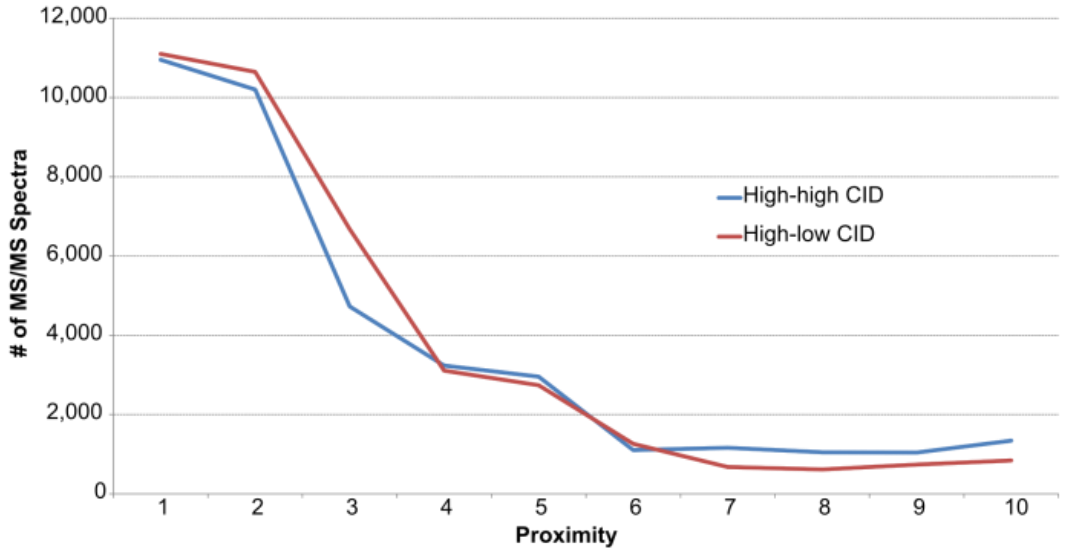


Figure 6.3. Illustration of site-determining b- and y-ions.

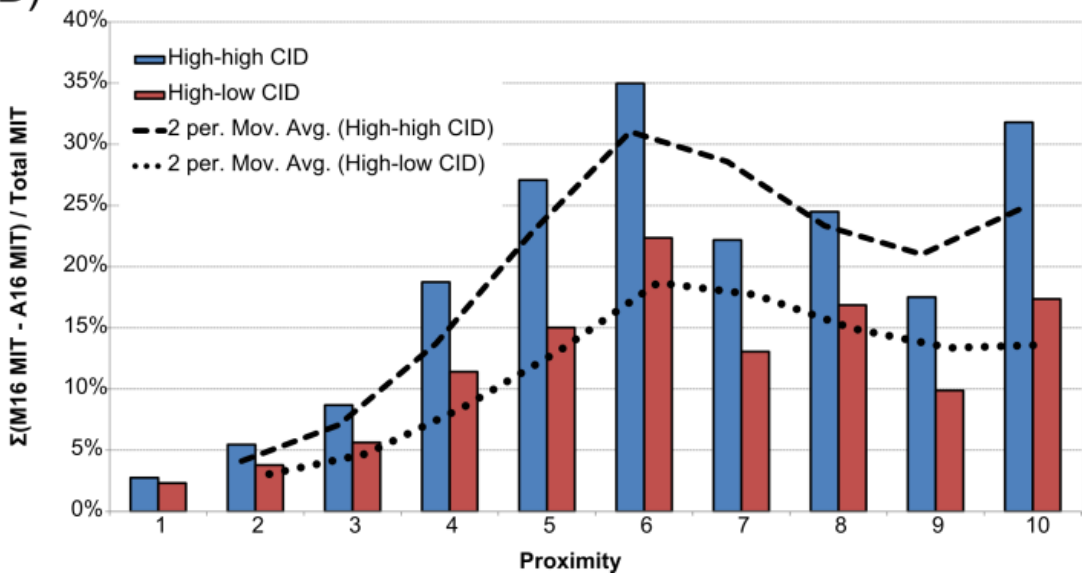
As anticipated, the number of CS declined as the distance between the methionine and alanine sites increased (Figure 6.4A). This observation is the result of an overall increase in the number of discriminatory b- and y-ions, which provides a more definitive spectral fingerprint. Also shown in this figure, the frequency of CS decreased at a similar rate for the two MS strategies. This was expected as both strategies perform collision-induced dissociation (CID); the MS/MS spectra will contain the same percentage of backbone fragmentation. Interestingly, both MS strategies show a clear inflection point when the proximity was ~6 amino acid residues. We suspect that this point represents the distance that provides the most discrimination between the two types of mass shifts, 1) those belonging to a methionine sulfoxide and 2) those more likely due to a SAAP. For distances greater than 6, the mass shift locations likely approach the terminal ends of the peptide sequence. In general, mass shifts located near the ends of a peptide sequence tend to be assigned less reliably than those near the center, which explains why a level of ambiguity remains. These observations are further corroborated by comparing the total matched ion intensity (MITs) of the b- and y-ion series for each peptide sequences that differed only by the location of a +16 Da mass shift. That is, for each ambiguous spectrum, we calculated the difference between the total MIT of the methionine (+16 Da) sequence and the total MIT of the alanine (+16 Da) sequence. Figure 6.4B shows the distribution of the percent difference between two potential sites for each distance. As shown, the maximum difference between the two theoretical mass shift sites occurred when the site locations were 6 amino acids apart. Although we suspected a high level of uncertainty for proximal sites, we demonstrated the likelihood of precise site localization is severely diminished when the number of site-determining b and y ions fall below 12. Notably, the vast majority of the CS (68% high-high and 70% high-low) belong to peptides containing two potential possibilities that are less than four amino acids apart. Cleary, these spectra have little or no site-determining information for proper site placement, which would be necessary for confident SAAP identification.

Figure 6.4. Identifying the level of ambiguity between adjacent mass shift sites. MS/MS spectra collected using a high-high (blue) and high-low strategy were interpreted by MyriMatch to identify all permissible +16 Da modifications on either alanine or methionine residues. Only contentious spectra (CS), MS/MS spectra that matched to the same peptide sequence but differed in the placement of the modification (i.e., at alanine or methionine), were plotted. (A) The frequency distribution of CS illustrates that level of ambiguity is strongly dependent on the distances between two potential modifications sites. (B) A matched ion intensity (MIT) was calculated for the two site positions and the difference between the matched ion intensity (MIT) values was calculated for each CS as a function of the proximity. A moving average trendline was provided for both the high-high (dashed-line) and high-low (dotted-line) strategy to highlight the earliest maximal difference in the matched ion intensities.

(A)


12,000
10,000
8,000
6,000
4,000
2,000
0
# of MS/MS Spectra
1
2
3
4
5
6
7
8
9
10
Proximity
High-high CID
High-low CID

(B)


40%
35%
30%
25%
20%
15%
10%
5%
0%
Σ(M16 MIT - A16 MIT) / Total MIT
1
2
3
4
5
6
7
8
9
10
Proximity
High-high CID
High-low CID
2 per. Mov. Avg. (High-high CID)
2 per. Mov. Avg. (High-low CID)

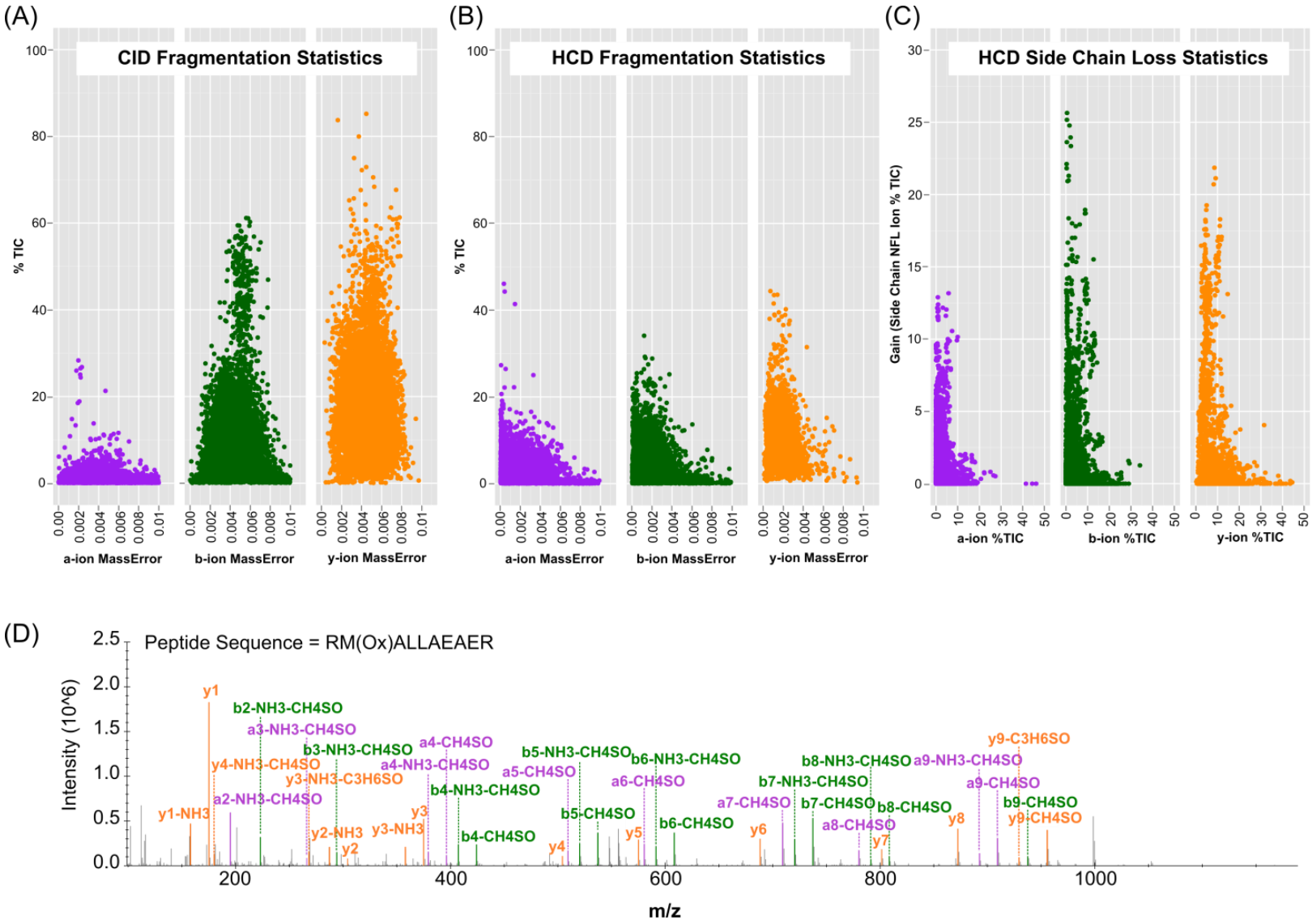
As others have shown, these observations highlight how precise site localization can be challenging for search algorithms when there are few site-determining fragment ions[261]. Presently, additional software is available to calculate the probability of correct localization for each site[262-265]. Though calculating a probability-based score provides a measure of certainty, spectra with insufficient site-determining ions (i.e., peptides with proximal residue sites and spectra featuring incomplete fragmentation) remain logistical problems. In other words, precise site localization in CID fragmentation spectra can be difficult when the distance between the two likely sites is less than 6 amino acids apart. Nevertheless, an immediate alternative approach is available to provide additional information for discriminating between SAAPs and what we suspect is the most common chemical modification mistaken for SAAPs: methionine oxidations (*vide infra*).

6.3.2 Identification of Methionine Sulfoxide Sites Using High Energy Dissociation (HCD)

For peptide-sequence tagging, we employed collision induced dissociation (CID), which is by far the most frequently used technique in proteomics for peptide sequencing. When CID fragmentation techniques are applied, the widely accepted model that describes the dissociation process designates b- and y-ion series as the most prevalent types[46-47]. The primary fragment ions and their contribution to the overall intensity coverage for a single CID run are illustrated in Figure 6.5A. In principle, complete coverage of the entire b- and y-ion series ions allows full annotation of the amino acid sequence of a peptide. As detailed in the section above, this information may be insufficient for definitively localizing mass shifts. However, there are alternative fragmentation processes that could benefit this task.

Introduced in 2007, higher energy collisional dissociation (HCD) fragmentation became available on the Orbitrap platforms[266]. In a dedicated collisional cell, peptide ions are subjected to a beam-type fragmentation process, where primary fragment ions retain kinetic energy and are therefore more likely to fragment again. In general, HCD ion types are expected to follow the fragmentation rules modeled from CID.

Figure 6.5. Fragmentations statistics of CID and HCD spectra. Only peptide spectrum matches (PSMs) meeting the following criteria were graphed: 1) PSMs identified by both CID and HCD strategies and 2) PSMs containing at least one methionine and a modified alanine (+16 Da) residue. A- (purple), b- (green), and y- (yellow) series were plotted. For each CID spectrum (A) and HCD spectrum (B), the percentage of the total ion current (TIC) attributable to a particular fragment ion series was plotted. (C) If a spectrum contained peaks which could unambiguously assigned as neutral losses from methionine sulfoxide, the additional intensity coverage for ambiguous spectra was calculated. (D) As an example, the HCD spectrum with the maximum additional coverage achievable (31%) was provided. Here, only the top 20 most abundant fragment ions were highlighted.


(A)
CID Fragmentation Statistics
% TIC
a-ion MassError
b-ion MassError
y-ion MassError
(B)
HCD Fragmentation Statistics
% TIC
a-ion MassError
b-ion MassError
y-ion MassError
(C)
HCD Side Chain Loss Statistics
Gain (Side Chain NFL Ion % TIC)
a-ion %TIC
b-ion %TIC
y-ion %TIC
(D)
Peptide Sequence = RM(Ox)ALLAEAER
Intensity (10^6)
m/z
y1
b2-NH3-CH4SO
a3-NH3-CH4SO
b3-NH3-CH4SO
a4-CH4SO
y4-NH3-CH4SO
a4-NH3-CH4SO
b5-NH3-CH4SO
b6-NH3-CH4SO
b8-NH3-CH4SO
a9-NH3-CH4SO
y9-C3H6SO
a5-CH4SO
a6-CH4SO
b7-NH3-CH4SO
a9-CH4SO
y3-NH3-C3H6SO
b4-NH3-CH4SO
a7-CH4SO
b7-CH4SO
b8-CH4SO
b9-CH4SO
a2-NH3-CH4SO
y1-NH3
y3
b5-CH4SO
y8
y9-CH4SO
y2-NH3
y3-NH3
b6-CH4SO
a8-CH4SO
b4-CH4SO
y5
y6
y7
y2
y4

Therefore, regular ions (b- and y-type ions) derived from backbone fragmentation are expected to be among the most abundant types observed. Besides a slightly lower contribution of the b- and y-ion series to the total TIC collected in each scan, the observed primary fragment ions and their overall intensities in a HCD run are comparable to CID (Figure 6.5B). A prominent difference, however, is larger contribution of the a-type ion series, which are derived from b-ions by losing CO. Moreover, as a direct consequence of the beam-type fragmentation process, the primary fragment ions are subjected to additional fragmentation pathways and consequently give rise to various ion types beyond those typically observed in CID[267]. A large portion of such ions are those involving neutral losses; the loss of water and ammonia are by far the most frequently observed. Another frequently observed class is the neutral loss of an amino acid side chain. In fact, the side chain of methionine sulfoxide is prone to cleavage[268], producing ions with a specific neutral fragment loss (NFL). Since search algorithms only consider backbone fragmentation (i.e., a-, b-, and y- ions) and some of their neutral losses (NH_3 and H_2O), a large percentage of the content in HCD spectra remain unassigned. Though many of these peaks belong to internal fragment ions and immonium ions, there are peaks which can be unambiguously assigned as neutral losses from methionine sulfoxide, based on the knowledge of how they fragment and the calculation of their fragment masses. Therefore, we exploited HCD fragmentation to identify the presence and precise location of methionine oxidations.

Again, the analysis was constrained to 'DENA' leaf samples and measurements were collected by the LTQ Orbitrap Pro mass spectrometer, which features improved sensitivity and HCD capability compared to its predecessors. HCD fragmentation was performed in the dedicated octopole collisional cell and fragment ions were detected in the Orbitrap. To test the suitability of this approach, the collected spectra were searched by MyriMatch using a directed method: alanine (+16 Da) was considered as the only dynamic modification. With this approach, the search algorithm considers the location of the mass shift irrespective of a neighboring methionine sites. Methionine was intentionally neglected during the peptide-spectrum matching process to eliminate the MyriMatch scoring system from the discrimination process. HCD spectra that matched a

peptide sequence containing a modified alanine (+16 Da) and at least one methionine were further interpreted. This step restricted the analysis to 4,943 spectra, which matched to 1,175 peptides. When annotating HCD peptide-spectrum matches, we looked for the presence of the characteristic neutral loss ions from the primary fragment ions (a, b, and y) of a peptide containing methionine sulfoxide (Figure 6.5D). As mentioned previously, the loss of water and ammonia from primary fragment ions are frequently observed. Therefore, these additional small molecule losses were taken into consideration when applicable.

For each spectrum, we calculated the percent gain in matched ion intensity when considering peaks attributable to the cleavage of a methionine sulfoxide side chain. Figure 6.5C depicts their overall contribution for each ion series: 96% of the spectra and 81% of the peptides exhibit at least one neutral loss from a methionine sulfoxide residue. With only a slight increase in the relative abundance of b-ions, the trends observed for each ion series (Figure 6.5C) agree with their expected contribution in a typical HCD run (Figure 6.5B). The most prominent fragmentation process observed was the neutral loss of methane sulfenic acid (CH_4SO). This chemical species exhibited a higher percentage of side change cleavage relative to the frequencies of the other fragment ions and could be observed in 83% of all MS/MS spectra exhibiting side chain loss. Despite only occurring when a fragment ion contains a methionine sulfoxide residue, i.e., CH_4SO, C_3H_6SO and C_3H_8SO, the three species could be found relatively abundant in the spectra, 3%, 1%, and 1% respectively. While their mean contribution to the overall intensity coverage was 5%, the maximum additional coverage achievable was 31% (Figure 6.5D). The gain in spectral information is promising: if searching algorithms could consider these characteristic permutations during the identification process, the false localization rate of oxidation events would be minimized. It should be noted, that the HCD fragmentation process is not only beneficial for the localization of methionine oxidations, but also for other modification events that have characteristic neutral losses, such as phosphorylations.

6.4 Identification of Sequence Variants by Integrating Genomics, Transcriptomics and Proteomics

As described in Chapter 3, SAAPs can be identified by exploiting expressed sequence tag libraries or whole-genome resequencing information by appending the genomic information to the reference proteome database. Despite the obvious potential to survey every molecular variation, this process often leads to an unacceptable increase in the database size, which directly leads to a higher risk of false positive identifications. In Chapter 6.3, the peptide sequencing tagging method - an effective hybrid approach (has elements of *de novo* and database searching algorithms) limited the search space by considering only variations that are extensions of peptides from a reference database - promised an unbiased look at the proteome's molecular phenotype; however, the ability to exclusively localize each polymorphism remains an ongoing challenge.

Amid recent advancements in RNA sequencing (RNA-Seq)[15], efforts have been made to integrate RNA-Seq data into proteomics studies[269-270]. The main advantage of using RNA-Seq data over whole-genome resequencing is the lower complexity (absence of introns and intergenic regions) and, in addition, RNA-Seq technology offers an opportunity to obtain transcript abundances as well as sequence variations concurrently. Recently, Wang *et al.* presented a sophisticated workflow that creates sample-specific protein sequence databases from RNA-Seq data to enable more variation-aware protein discovery investigations[271]. To evaluate this method, we applied a three-step strategy: (1) map RNA sequencing data onto the *Populus* proteome and identify SAAPs, (2) verify sequence variants with bottom-up proteomics, and (3) validate identifications with whole-genome resequencing data. For this investigation, we integrated results of (a) whole-genome resequencing, (b) RNA sequencing, and (c) MS/MS protein sequencing from stem tissue of the *P. trichocarpa* genotype 'DENA'.

As the first step, high-throughput RNA-Seq data (Illumina; 3 replicates; 50bp paired-end)[118] was aligned[272] to the *P. trichocarpa* reference genome (v3; 73,013 transcripts/proteins). Table 6.6 provides a summary of the mapping results.

Table 6.6. Summary of RNA-seq mapping results.

Accession	Read (50*2)	Total reads	Read1 mapped	Read2 mapped	Proper mapped
SRR072980	10,352,888		27,375,114	27,023,456	22,503,696
SRR072990	10,635,737	62,539,176 (31,269,588)	22,822,720 (72.99%)	22,642,943 (72.41%)	20,038,518(64.08%)
SRR073110	10,280,963		21,415,226 (68.49%)	21,267,965 (68.01%)	19,016,488(60.81%)

Among the 31,269,588 reads, 72% were mapped to the genome. Next, SAMtools[273] were used to call sequence variations. To reduce the number of false positives, the data was filtered based on the following criteria: (i) SNP quality above 20, (ii) mapping quality above 25, (iii) coverage above 6, and (iv) an alternative base must be supported by at least 3 reads. Table 6.7 shows the total number of single nucleotide variations (SNVs) as well as the SNVs locations. Generally speaking, the majority (93.61%) of SNVs were located in exon regions. According to the comparison of RPKM[274] distribution for exon and intron regions (i.e., an estimate for the absolute expression levels of transcripts), we used an RPKM threshold of 1 to remove the transcripts with low expression level – that is, those detected transcripts with RPKM values less than 1 are likely background noise. Nonsynonymous SNVs (i.e., SAAPs) which were located in the coding region were introduced into the protein sequences. Using the above criteria, we constructed a customized proteome sequence database (63, 241 total sequences; 46,423 RPKM >1; 15,278 SAAPs), which was used to search the MS/MS data collected from the 'DENA' stem sample. In brief, all MS/MS spectra were searched with the MyriMatch algorithm against the customized database. Following the search, peptide identifications were assembled into proteins and filtered by IDPicker to achieve a FDR of <2% (PSM-level).

In order to compare the results across all three levels of data (genome::transcriptome::proteome), each protein/transcript identified was referenced by its gene name, which effectively reduced the complexity of the analysis - a single gene can have multiple transcript names (i.e. protein names) as a result of alternative splicing. In order to localize each genetic variation to only a single location in the genome, we required that (1) each peptide sequence to be unique to a particular gene and (2) each peptide sequence only occur once in a particular protein/gene. Using these filters, we identified 313 SAAPs, which mapped to 284 genes. Since whole-genome sequencing data has been made available for 'DENA'[114], we were able to provide an additional level of validation. Of the 313 SAAPs that were confidently identified, 185 (~60%) were also identified by whole-genome resequencing.

Table 6.7. Gene location of single nucleotide variants.

Total	3'UTR	5'UTR	Coding	Intergenic	Intron
67,050	16,478(24.58%)	6,425(9.58%)	39,863(59.45%)	2,185(3.26%)	2,099(3.13%)

We suspect that this discrepancy is likely an artifact of insufficient sequencing depth in the whole-genome resequencing data. To explain away this discrepancy, we are currently assessing the level of sequencing depth for each SAAP chromosome position across each data set. While additional analysis regarding the biological relevance of these genes has not yet been performed, we suspect that further characterization may reveal that these genes are involved in biological processes responsible for adaptive physiological responses.

Since *Populus* is a diploid organism (i.e., has two matching sets of chromosomes), each gene will have a pair of alleles and, in general, alleles vary in their degree of zygosity (i.e., the degree of their genetic variation). Therefore, for each SAAP call identified, we determined whether the position was homozygous (i.e., identical) or heterozygous (i.e., two alternative variants). For zygosity determination at each location, we used the maximum likelihood estimator of the allele frequency (assuming Hardy-Weinberg expectations[275]). Out of the 185 SAAPs, 105 positions were homozygous and 80 were heterozygous. For 37 of the heterozygous locations, we identified both versions of the peptide (Table 6.8).

To investigate the potential impact of the identified SAAPs, we used the SNAP (screening for non-acceptable polymorphisms)[276] method to calculate whether the amino acid substitution are predicted neutral (i.e., no effect) or non-neutral mutations. The calculation is based on protein sequence-based predictions of solvent accessibility and secondary structure from PROF[277], flexibility from PROFbval[278], function effects from SIFT[279], as well as conservation information from PSI-BLAST[280], PSIC[281] and PFAM[232] annotations. As an output, SNAP reports three values: (1) a binary prediction (neutral/non-neutral), (2) a reliability index (RI; range 0 to 9), and (3) the expected accuracy. Overall, only 3 of the heterozygous alleles had a sequence variation that was considered non-neutral.

Table 6.8. Peptides identified from heterozygous genes.

#	Gene Name	Type	Variation	Allele A SpC	Allele B SpC	Status	RI	Accuracy
1	Potri.010G028600	Hetero | Ref	A240T	20	9	Non-neutral	1	63%
2	Potri.010G114200	Hetero | Ref	S403G	3	6	Neutral	5	89%
3	Potri.010G142600	Hetero | Ref	T16N	57	36	Neutral	5	89%
4	Potri.010G127500	Hetero | Ref	I203V	11	68	Neutral	4	85%
5	Potri.009G155200	Hetero | Ref	I429L	4	4	Neutral	8	96%
6	Potri.009G020800	Hetero | Ref	G234E	1	6	Neutral	6	92%
7	Potri.009G094800	Hetero | Ref	V2A	4	5	Neutral	8	96%
8	Potri.009G080200	Hetero | Ref	S228N	4	1	Neutral	7	94%
9	Potri.004G007900	Hetero | Alt	E172Q	1	9	Neutral	0	53%
10	Potri.004G067000	Hetero | Ref	I40L	3	3	Neutral	3	78%
11	Potri.004G010100	Hetero | Ref	I180V	10	27	Neutral	2	69%
12	Potri.013G058300	Hetero | Ref	P251L	7	9	Neutral	5	89%
13	Potri.013G032300	Hetero | Ref	T84M	1	3	Neutral	5	89%
14	Potri.013G006100	Hetero | Ref	I372L	21	21	Neutral	7	94%
15	Potri.013G079000	Hetero | Ref	S70T	7	7	Neutral	7	94%
16	Potri.006G267500	Hetero | Ref	V81I	6	1	Neutral	6	92%
17	Potri.006G224500	Hetero | Ref	L3Q	2	6	Neutral	8	96%
18	Potri.016G119800	Hetero | Ref	S171T	8	4	Neutral	3	78%
19	Potri.016G013000	Hetero | Ref	K107E	15	12	Neutral	7	94%
20	Potri.016G013000	Hetero | Ref	E108K	26	12	Neutral	6	92%
21	Potri.016G091100	Hetero | Ref	E338D	6	2	Neutral	6	92%
22	Potri.015G070600	Hetero | Ref	P673	7	4	Neutral	8	96%
23	Potri.002G166800	Hetero | Ref	V67A	15	5	Neutral	4	85%
24	Potri.002G184300	Hetero | Ref	D293G	6	1	Neutral	3	78%
25	Potri.019G088100	Hetero | Ref	A68S	13	20	Neutral	7	94%
26	Potri.005G032800	Hetero | Ref	A216T	5	2	Neutral	4	85%
27	Potri.014G073000	Hetero | Ref	E312D	2	3	Neutral	7	94%
28	Potri.014G160000	Hetero | Ref	I245V	1	11	Neutral	7	94%
29	Potri.014G139500	Hetero | Ref	P714L	27	28	Neutral	2	69%
30	Potri.008G117600	Hetero | Ref	Y164S	2	8	Neutral	2	69%
31	Potri.008G107900	Hetero | Ref	F246L	1	4	Neutral	6	92%
32	Potri.001G358300	Hetero | Ref	I284T	19	69	Non-neutral	4	82%
33	Potri.001G198000	Hetero | Ref	A138S	6	13	Neutral	7	94%
34	Potri.001G214500	Hetero | Ref	L278I	17	17	Non-neutral	1	63%
35	Potri.001G278900	Hetero | Ref	K178N	9	12	Neutral	3	78%
36	Potri.001G018200	Hetero | Ref	N543I	20	1	Neutral	0	53%
37	Potri.012G095500	Hetero | Ref	M460I	2	1	Neutral	7	94%

* There are two types of heterozygous peptides: “Hetero | Ref” (i.e., peptide variants that match the reference genome) and “Hetero | Alt” (i.e., peptide variants that are specific to the DENA genotype). A spectra count (SpC) was provided for both alleles. SNAP results were obtained for each variation. The status (non-neutral or neutral), the reliability index (RI), and the accuracy.

Interestingly, two peptide variants of the protein (Potri.001G358300.1) had almost a four-fold difference in their total number of spectra count (19 SpC versus 69 SpC), suggesting that the two protein isoforms have different allele frequencies. In an attempt to elucidate a functional role, we searched the protein sequence against the Pfam database. By investigating protein family membership, two functional domains could be identified: RSN1_TM (amino acid residues 5 through 165) and DUF221 (amino acid residues 314 through 636). The RSN1_TM family often represents the first three transmembrane (TM) regions of a 9-11TM protein that are associated with late exocytosis and is associated with Golgi transport of vesicles to the cell surface[282]- a transmembrane prediction using Hidden Markov Models (TMHMM) analysis revealed that the protein has 9 TM-helices (Table 6.9). Unfortunately, the DUF221 domain has no known function. To examine whether the function of the protein has been identified in another organism, we searched the protein against the BLASTP database. One of the top hits was an *A. thaliana* protein with 67% sequence similarity which has been annotated as early-responsive to dehydration 4 (ERD4). In general, osmotic homeostasis is maintained in a cell through continuous cycles of rehydration-dehydration, where water and solute flux is under equilibrium. When a cell becomes dehydrated, it becomes depolarized. As a stress response, the cell must target efflux protein pumps to the plasma membrane in order to polarize the cell, reducing the internal concentration of solute which, in turn, allows water enter the cell (i.e., osmosis). Interestingly, it has been shown that RSN1-TM domains are found in proteins required for maintaining ion homeostasis in cells. In one particular study, the protein played an important role in maintaining cellular polarity by sorting plasma membrane proteins, targeting a sodium pumping ATPase to the cell surface[282]. Due to their structural similarity, it is logical to assume that the *Populus* protein (Potri.001G358300.1) plays an important role in the interaction of secretory vesicles with the plasma membrane in order for the cell to maintain cellular homeostasis.

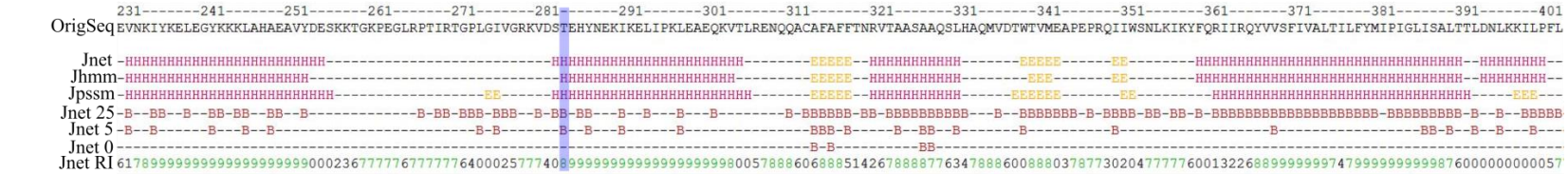
Because structure likely plays an important role in its targeting mechanisms, we assessed the location of the SAAP and how that might impact the secondary structure of the protein.

Table 6.9. Transmembrane helix prediction results for Potri.001G358300.1.

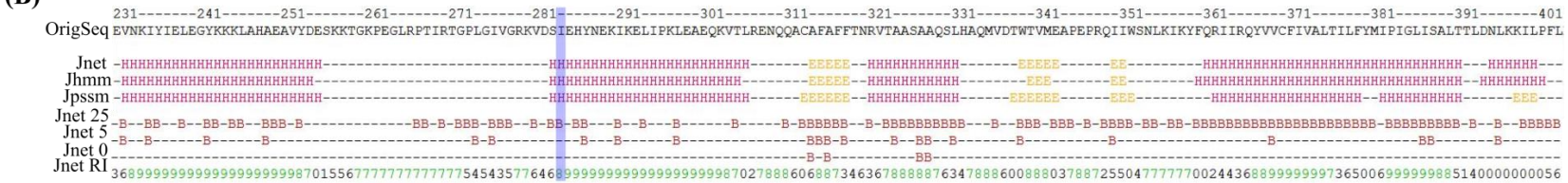
TMHMM Result	Start	End
Outside	1	4
TMhelix	5	27
Inside	28	84
TMhelix	85	107
Outside	108	144
TMhelix	145	164
Inside	165	367
TMhelix	368	390
Outside	391	458
TMhelix	459	481
Inside	482	489
TMhelix	490	512
Outside	513	563
TMhelix	564	595
Inside	596	615
TMhelix	616	638
Outside	639	642
TMhelix	643	662
Inside	663	724

Figure 6.6. Illustration of Jpred results for the heterozygous peptides for protein Potri.001G358300.1. Here, a four turn alpha helix is represented as a red “H”, an extended strand in a beta-sheet conformation is represented as a yellow “E”, and a buried amino acid residue in a secondary structure is represented by a red “B”. The blue box highlights the amino acid position under investigation. (A) Allele A contains a threonine amino acid residue at position 284. (B) Allele B contains an isoleucine amino acid residue at position 284. A portion of the original sequence (OrigSeq) is provided as the following Jpred results: Jnet (the final secondary structure prediction), Jhmm (Jnet Hidden-Markov model prediction), Jpssm (Jnet PSIBLAST position-specific scoring matrix profile prediction), Jnet 25, 5, and 0 (Jnet prediction of burial, less than 25%, 5%, and 0% solvent accessibility), and Jnet Rel (Jnet reliability of prediction accuracy, ranges from 0 to 9, larger is better).

(A)


231 241 251 261 271 281 291 301 311 321 331 341 351 361 371 381 391 401
OrigSeq EVNKIYKELEGYKKKLAHAEAVYDESKKTGKPEGLRPTIRTGPLGIVGRKVDSTEHYNEKIKELIPKLEAEQKVTLRENQQACAFAFFTNRVTAASAAQSLHAQMVDTWTVMEAPEPRQIIWSNLKIKYFQRIIRQYVVSFIVALTILFYMIPIGLISALTTLDNLKKILPFL
Jnet
Jhmm
Jpssm
Jnet 25
Jnet 5
Jnet 0
Jnet RI

(B)


231 241 251 261 271 281 291 301 311 321 331 341 351 361 371 381 391 401
OrigSeq EVNKIYIELEGYKKKLAHAEAVYDESKKTGKPEGLRPTIRTGPLGIVGRKVDSIEHYNEKIKELIPKLEAEQKVTLRENQQACAFAFFTNRVTAASAAQSLHAQMVDTWTVMEAPEPRQIIWSNLKIKYFQRIIRQYVVCFIVALTILFYMIPIGLISALTTLDNLKKILPFL
Jnet
Jhmm
Jpssm
Jnet 25
Jnet 5
Jnet 0
Jnet RI

To see how the non-neutral SAAP (T284) might impact protein structure, we submitted the two protein sequences to the JPRED 3 secondary structure prediction server[283]. For discussion purposes, the two proteins will be referred to as Allele A (carries the non-neutral mutation) and Allele B. As shown in Figure 6.6, the SAAP occurs at a highly conserved location, which is at the second position of a predicted alpha helix. As illustrated in the TMHMM results, this specific helix is located on the inside of the cell. The substitution from an isoleucine to a threonine is predicted to impact the secondary structure of the protein by not only changing the physiochemical properties of the helix it's located on, but also upstream and downstream secondary structures (Figure 6B). Although further analysis is required, we suspect that these structural changes likely alter how Allele B interacts with its target protein(s) and, since Allele B was more abundant, this is likely a favorable substitution.

6.5 Conclusions

Here we implemented automated sequence tag inferences (DirecTag) and reconciliation (TagRecon) for the identification of unanticipated sequence variants at a global level. Together, careful search space selection and the availability of high mass-accuracy data maximized the sensitivity of the experiment and improved the overall integrity of the data set. The large-scale study yielded a broad and quantitative view of the frequencies and abundances of various single amino acid polymorphisms in *Populus*. Despite the limited sequence coverage afforded in a typical bottom-up approach, peptide sequence variants were nearly observed in every protein. Overall, we were able to generate a dataset containing 6,653 attested peptide sequence variants.

Though we have shown the potential of peptide sequencing tagging in *Populus*, a high-throughput and automated assignment of mass shifts to the correct amino acid remains a challenge for these large-scale studies. A widely acknowledged problem, precise site localization becomes difficult when multiple residues within a single peptide can be modified. When the distance between two potential sites decreases, the theoretical fragmentation peaks of the two candidate annotations become more similar. As a result, there are fewer site-determining ions available to uniquely assign a mass shift to a

specific residue. Especially since the CID fragmentation process is often incomplete, the identification of a full series of the b- or y-ion type is rarely achieved. Although search algorithms report the highest scoring modified peptide, insufficient site-determining ions may lead to the incorrect localization of mass shifts. These shortcomings were clearly apparent in our study, as the most abundant chemical modification could masquerade a sequence variant (Ala→Ser). Currently, there are more sensitive approaches available, such as the ASCORE method, which calculate a probability-based for specific site locations. Although these scores can generally discriminate alternative sites, a smaller spacing of the two sites within a peptide sequences can lower the performance of the scores. Within our data set, the spacing of alternative sites within a peptide greatly influenced localization of the mass shift: the maximum discriminating evidence did not occur until two alternative oxidation sites were 6 amino acids apart. More importantly, peptides with potential sites <4 amino acids apart (~70% of all the CS) had insufficient evidence for confident site placement.

Owing to the frequency of methionine oxidations, we exploited HCD fragmentation to assess their location objectively without the need for using probability-based scores. Because the HCD fragmentation behavior of methionine sulfoxide containing peptides can be quite distinct, we were able to empirically collect information that facilitated the localization of ambiguous +16 Da mass shifts. In contrast to CID spectra, in which only the regular ion series (a, b, and y) are available, HCD spectra contain characteristic neutral fragment loss ions which enabled the explicit identification of methionine sulfoxide residues. If search algorithms could make use of the available additional spectral information, we suspect that HCD will enable improved site placement for de novo sequencing and hybrid peptide sequencing tagging-based approaches.

In addition to the PST approach, we implemented an integrated omics-based approach to identify SAAPs. Collectively, the integration of genomics, transcriptomics, and proteomics provided a more confident set of SAAP identifications. In general, we are hopeful that the data obtained can be used to help advance the peptide sequence tagging algorithms or, more broadly, *de novo* algorithms. One advantage of using the bottom-up

approach for SAAP characterization is that we can now further interrogate the identified allelic frequencies at the protein level and, perhaps, quantitate the relative abundance of both protein isoforms. This approach, for instance, could be used to detect the relative excess or deficit of certain allelic frequencies across a natural population, providing information that lead to the discovery of a novel protein form that is responsible for a favorable phenotype. Furthermore, rather than just focusing on SAAPs, this framework could be adjusted for the identification of novel alternative splice forms as well as insertion/deletion events. Currently, we are expanding the scale and scope of the integrated omics-based analysis to a second *P. trichocarpa genotype* as well as other tissue types.

CHAPTER 7

FUTURE OUTLOOK, REMAINING CHALLENGES, AND CONCLUSIONS

7.1 Overview

The ability to identify and quantify proteins with high precision in *Populus* samples is an essential requirement in bioenergy research. Although mass spectrometry-based proteomics has emerged as a promising technique for such precise characterizations, its full potential had not been realized for *Populus*. In Chapter 1, three specific goals were outlined: (1) design a bioinformatic workflow that addresses a protein inference problem with respect to genetic duplications develop methods to maximize protein identification, (2) develop a method to maximize protein identifications, and (3) develop a mass spectrometry-based method for the identification of single amino acid polymorphisms. Described in Chapter 3, a bioinformatics approach was developed to better define the measured proteome. By grouping proteins together by sequence similarity, proteome data sets were improved with respect to protein inference because they are richer in relevant information and omit information on shared peptides that could not be conclusively mapped to the proteome. The ability to accurately compile and organize the collection of data into a proteome map better defined the achievable coverage. A method to maximize proteome coverage in the genus *Populus* was described in Chapter 4 and implemented in Chapter 5. The development of enhanced protein retrieval through a detergent-based lysis approach and maximized peptide sampling via the dual-pressure linear ion trap mass spectrometer (LTQ Velos) facilitated the identification of 25% of the predicted proteome space. The technological advancements, specifically spectral-acquisition and sequencing speed, afforded the deepest look into the *Populus* proteome, with protein abundances spanning 6 orders of magnitude. In the emerging systems biology paradigm, there is a strong emphasis on determining the context of proteins within a biological system. Therefore, the ability to comprehensively

identify proteins was exploited to create a *Populus* proteome atlas, representing a visual collection of the data by highlighting relationships of the spatial arrangements of proteins as well as their abundance. The proteome atlas offered varying proteome perspectives that depicted a complex wiring of biochemical reactions with varying resolution. Emulating developments in genomics and transcriptomics, two methods were developed to decode the plasticity of the *Populus* proteome. In order to detect polymorphic deviations from a static reference genome, two different approaches were made available. Without *a priori* knowledge of sequence variants, a peptide sequence tagging (PST) approach was developed and implemented to reveal a large number of unexpected SAAPs that would have otherwise been missed by a traditional database search algorithm. With parallel availability of genomic and transcriptomic information, an integrated omics-based approach becomes feasible. Described in Chapter 6, an approached was developed to combine information from genomic, transcriptomic, and proteomic data sets to confidently identify single amino acid polymorphisms. By enabling the identification of single amino acid polymorphisms at the protein level, these two methods will improve our understanding of natural variation across the genus *Populus* and provide a useful approach to characterize the mechanisms underpinning target traits. Altogether, the developments described in Chapters 3-5 will greatly benefit the characterization of the genus *Populus* and potentially other eukaryotic organisms as well. Chapter 6 outlines how the developments provide a springboard for future studies. For *Populus* proteomics, this means that the emphasis will shift from the continued discovery of proteome maps towards determining the relevant biological information about proteins in the context of a biological system. In addition, this chapter outlines current challenges as well as a discussion of what has accomplished in this dissertation and how that impacts life science in general.

7.2 Status of *Populus* Proteome Characterization

As discussed in Chapter 3, utilizing protein grouping as an approach to assemble peptide sequences into proteins afforded a straightforward and accurate way to report the proteins present in each *Populus* sample. Although protein grouping was initially

designed for the *Populus* proteome, data acquired from human as well as microbial community samples will also require a similar, if not the same, approach. Like *Populus*, a large percentage of the proteins expressed by human cells are homologous proteins or splicing variants (see UniProt website). Therefore, the bioinformatic approach described in this dissertation could be readily applied to future studies involving human samples. Since prokaryotic organisms have highly reduced genomes with very little sequence redundancy, one would not expect a population of prokaryotic cells to generate a large number of shared peptides. However, when dealing with microbial communities, peptides can map to similar proteins from different strains of a particular organism or, for diverse communities, peptide will likely map to highly conserved protein motifs. For example, the human gut microbiota is an assortment of a number of bacteria, archaea, viruses, and unicellular eukaryotes. In fact, a recent study suggests that the human gut microbiota consists of over 1,000 bacterial species. Although protein grouping by sequence homology can be implemented, this may eliminate any strain-resolving resolution. Also within the collection of microorganisms, there are microorganisms that perform similar roles - for example, there is methanogenic microbiota which produces methane as an end-product of amino acid metabolism – and therefore it is likely that these proteins share similar functional domains. In this scenario, a slightly different protein inference approach may be required because, rather than entire protein sequences sharing homology, only a small percentage of each protein will be shared. That is, protein grouping by sequence homology may not be the best approach because this would require sequence similarity thresholds to be significantly lower (i.e., near 50%), which would lower the resolution to protein families (i.e., only a functional perspective would be available).

Five years ago, an extensive LC-MS/MS analysis (60 MS/MS runs) on stem tissue culminated in the identification of ~7,500 protein sequences. Today, one can readily identify ~10,000 protein sequences in a single LC-MS/MS when the established experimental framework is applied – the cumulative result of method and technical developments discussed in Chapter 4 and implemented in Chapter 5. With this established framework available, future studies should perform comparative analysis of

numerous *Populus* molecular phenotypes. Specific to the bioenergy research effort, future experiments should generate proteome maps for trees containing favorable traits such as disease resistance, drought tolerance, increased photosynthesis, and greater carbon allocations to stem diameter versus height growth. Analyzing global proteome changes could provide important information regarding the biological processes and pathways that are involved in establishing alternate phenotypes. Also highlighted in Chapter 5, another key advantage of bottom-up proteomics for systems biology is the capability of quantifying proteins. In general, the semi-quantitative label-free approach could provide indispensable information that guides the development of more targeted approaches. For example, the information gleamed from a discovery-based approach could be used to create an LC-MS/MS that only focuses on a particular pathway of interest. In this scenario, a list of known peptide precursor ion *m/z* values could be used to construct a data-dependent inclusion list. Rather than dynamically excluding *m/z* values during the course of a LC-MS run, the analyzer only triggers fragmentation for ions within the user-specified inclusion list. Using this approach, future studies could rapidly compare, with higher sensitivity, the relative abundance of pathways across multiple sample sets. Within the scope of bioenergy research, the approach should be applied to profile *Populus* and, eventually, switchgrass reduced-recalcitrance lines that have been identified through association genetics. Although the method and technological advancements were geared towards plant samples, similar versions of the experimental strategy have been implemented for other sample-types, such as the human gut microbiome.

Since SAAPs are expected to be a major cause of bioenergy-related phenotypes, the methods described in Chapter 6 will likely be applied in future studies to detect and quantify polymorphisms that affect the structure and/or function of a protein(s). In more broad applications, the use of the methods described in this dissertation should become more widespread in cancer-related research. In this case, proteomics data could reveal unexpected sequence variants that cause proteins to acquire oncogenic properties by, such as increase protein expression or a gain of function. In addition, the ability to generate RNA sequencing-derived proteome database should not only facilitate the identification of single nucleotide variants, but novel alternative splice forms as well. In fact, as the

technologies improve and become more widely available, it becomes viable to integrate some level of RNA-seq data with all future proteome data sets. Furthermore, RNA-seq derive proteome database will likely become the mainstay for studies with incomplete reference protein sequence databases (for example, microbial communities).

7.3 Status of Experimental Strategies and Remaining Challenges

Although a method for protein extraction has been demonstrated in this dissertation, the complexity of the resulting peptide mixtures remains problematic, necessitating further developments. With the current the experimental design, there are likely over 10,000 different protein isoforms that are being extracted and processed as a single unit. After digestion with trypsin, the highly complex peptide mixtures are separated by chromatography. Conservative estimates of the number of peptides present in standard MS analysis of complex cellular lysates suggests that there are likely over 100,000 detectable peptides available for identification. At present, developmental efforts have so far mainly concentrated on improving the separation of peptides prior to analysis and MS parameters to improve the identification rate of peptides (*vide infra*). However, it would be interesting to investigate how protein fractionation can improve proteome coverage. To date, the most widely applied protein fractionation is a gel-based approach (i.e., 1D-PAGE), in which proteins are separated by their size (i.e., molecular weight). One of the advantages of separating proteomes by size is that they fraction in a predicable fashion, such that one can select the number of fractions collected. In fact, such predictability can permit the isolation of proteins, allowing the enrichment of specific classes of protein isoforms. Unfortunately, there are many known disadvantages of using gel-based approaches - the most serious limitation is protein recovery. As an alternative to gel-based approaches, proteins are commonly fractionated by ultracentrifugation strategies. However, these approaches are inherently labor intensive and offer poor degree of resolution and recovery. Recently, improved proteome coverage has been demonstrated using gel-eluted liquid fractionation entrapment electrophoresis (GELFrEE). First described by Alan Doucette and colleagues, the new fractionation scheme provides the ability to perform molecular weight-based fractionation with liquid

phase recovery. The typical GELFrEE system consists of a continuous elution column, where proteins are electrophoretically eluted (i.e., eluted) from the end of the column and subsequently trapped in a collection chamber. One of the main benefits of the gel-free fractionation system is that it is immediately compatible with the current sample preparation procedure. In addition, the gel-free procedure partitions the complex protein mixtures into user-selectable molecular weight fractions. Using this device, *Populus* protein mixtures could be fractionated (number of fractions is yet to be determined) in order to reduce the complexity of the analyzed peptide mixtures. Although this would increase the number of samples that need to be performed per sample, the potential gain in dynamic range and the overall number of protein identifications would arguably outweigh this disadvantage.

Another challenge is sequencing as many peptides as possible in a single LC-MS/MS run. Even with more sophisticated sample preparations and chromatography, emphasis must be placed on optimizing data-dependent acquisition parameters. Despite technical advances, the entire population of peptides being analyzed at a particular time exceeds the threshold of MS/MS peak picking. Therefore, data-dependent acquisition parameters must be optimized to ensure that time is not wasted collected unusable fragmentation information. Currently, when using data-dependent acquisition on state-of-the-art instrumentation, roughly 300,000 MS/MS spectra can be collected per MS analysis. Out of those collected, only around 50% of those spectra result in the identification of a peptide sequence. Though many of the unassigned MS/MS spectra are not identified because their corresponding peptide is not investigated (for example, post-translational modifications; *vide infra*), many spectra remain unassigned because they are low-quality spectra (i.e., high-signal to noise and poor representation of fragment ions). To improve data acquisition, one must sample each detectable peptide at their highest signal in order to obtain the highest quality MS/MS spectra, while avoiding the collection of low-quality spectra. Since many low-quality spectra belong to identifiable peptide sequences, a data-dependent method should be designed to ensure that peptide fragmentation occurs when the signal of a peptide is at its highest point (i.e., the chromatographic peak apex), resulting in more high-quality spectra. When designing this

approach, one could optimize the available dynamic exclusion parameters. Specifically, emphasis should be placed on the number of repeat counts and the duration in between repeat analysis. Since the average elution time of a peptide is ~1 minute, a repeat count of 2 and a repeat duration of 30 seconds may prove the most optimal setting. In order to avoid the collection of low-quality spectra entirely, the MS analyzer could be parameterized to trigger fragmentation only on isotopic packets with an identifiable charge state or when an ion signal is above an acceptable threshold. Currently, fragmentation occurs regardless of charge state determination. If fragmentation was restricted to only precursor ions with a discernible charge state, not only would parts per million (ppm) mass tolerances become applicable, but also the quality of the MS/MS spectra would be higher. Moreover, the exclusion of +1 charge states should be enforced, since they contain less fragmentation information. Although some form of these optimizations currently exists in the proteomics field, they have yet to be optimized and tested for *Populus* samples. Even with advanced inclusion and exclusion features for peak selection, the identification of all detectable peptide sequences during an LC-MS/MS analysis still requires improvements in sequencing speed, sensitivity, and higher resolution precursor selection and isolation.

Though a method has been described in this dissertation for the identification of SAAPs, *Populus* MS/MS data remains to be searched for unexpected post-translational modifications. Of particular interest is the identification of glycosylation events. Glycosylation is considered one of the most important and most common form of post-translation modifications in plants, yet it is also considered the most difficult to analyze. Today, considerable work has been done to characterize the glycoproteome (i.e., the global analysis of glycoproteins) for many organisms; however, zero reports are available that detail a systematic screening of any plant glycoproteome. Clearly, this is a remarkable deficiency in the proteomics field. Though there are several challenges to overcome, an experimental strategy should be designed to characterize the *Populus* glycoproteome. Today, the major challenge of comprehensive glycoproteomic analysis arises from the variety of carbohydrates that can be linked to a protein. Though MS/MS fragmentation can be applied to identify the composition and structure of these

modifications, it is currently not feasible to comprehensively characterize these modifications in a global scale. Therefore, a glycoproteome analysis should focus on identifying which proteins are modified as well as the location of each modification. To accomplish this, an experimental strategy should include glycoprotein enrichment, a deglycosylation step and tandem mass spectrometry, followed by bioinformatic interpretation using the approach outlined in Chapter 6. In addition to glycosylation, the approach described in Chapter 6 should also be used to broadly identify other unexpected PTMs (i.e., methylation, acetylation, ubiquitination, phosphorylations, etc.). Like SAAP identifications, the precise localization of a PTM will be the most challenging task.

7.4 Concluding Perspective

With tremendous foresight, Nobel Prize Laureate J.J. Thomson observed in his book “Rays of Positive Electricity and Their Application to Chemical Analysis" that a new technique would be highly profitable to chemistry and solve many problems with “far greater ease” than by any other technique used to analyze chemicals. Over a century ago, J.J. Thomson developed the first mass spectrometer in which he observed that lighter atoms behaved differently than heavier atoms in a cathode ray tube. By measuring one physical property, mass, his fundamental measurement laid the foundation for scientific breakthroughs, which began with the discovery of a number of isotopes, their relative abundance, and their exact masses. As stated in Chapter 1, mass spectrometers constitute a large, very diverse, and widely applied family of analytical platforms that are used to identify unknown compounds, quantify known compounds, and elucidate the structure and chemical properties of simple molecules as well as complex macromolecules such as proteins. At the intersection of mass spectrometry and molecular biology, proteomics has become an indispensable field of study that continues to play an essential role in connecting genotypes to molecular phenotypes.

Although mass spectrometry-based proteomics has become more accessible to scientists with varying levels of expertise, the high cost of state-of-the-art equipment, which is in constant transformation, has restricted most high-impact proteomic analyses to a relatively small number of laboratories. Therefore, even though highly confident

identification of 10,000 proteins has become feasible in *Populus*, many studies still report only a 1,000 or fewer highly abundant protein species because they are limited by the available technology. No doubt, until this technological gap is reduced, the developmental pace of the field will remain relatively slow when compared to genomics or transcriptomics. The main reason for this discrepancy is that, unlike the sequencing of a genome and transcriptome, which conceptually and technically have become somewhat streamlined, there is no "kit" or protocol that is amenable to all types of proteomes. Likewise, there is currently no sequencing platform and computational infrastructure that can be used to rapidly sequence and assemble proteomes. As demonstrated in this dissertation, MS-based approaches and instrumentation are constantly being developed transformed in order to match the proteome under investigation.

Though the methods in this dissertation were developed and applied specifically for the genus *Populus*, I predict that some form of these methods will be applied to other species as well, such as humans. Due to the emerging field of systems biology, these powerful qualitative and quantitative methods will, without a doubt, be a critical asset in linking the physical and functional interactions of proteins to the dynamic networks that ultimately determine phenotypes. As a result, most scientists will gravitate away from the analysis of a small set of proteins towards comprehensive and reproducible large high-quality data sets that are being made available via MS-based proteomics. In this paradigm shift, there should be less emphasis on myopic investigations that focus only on a few protein "favorites" and more emphasis on discovery investigations that explore and broaden the scope of the measurable proteome. With the ever increasing technological advancements, these types of explorations should enable the detection and quantification of every protein within a defined proteome. Of course, it will be equally important to continue developing effective computational frameworks for the integration the knowledge made available at each omics-level. In addition, better integration of proteomics data with phenomic information will be necessary to not only construct biochemical networks, but also how these networks are perturbed by various endogenous or exogenous cues – an essential step in bridging a genotype and phenotype.

LIST OF REFERENCES

1. Gurdon, J. B.; Elsdale, T. R.; Fischberg, M., Sexually Mature Individuals of Xenopus-Laevis from the Transplantation of Single Somatic Nuclei. *Nature* **1958,** *182* (4627), 64-65.

2. Okita, K.; Ichisaka, T.; Yamanaka, S., Generation of germline-competent induced pluripotent stem cells. *Nature* **2007,** *448* (7151), 313-7.

3. Crick, F., Central dogma of molecular biology. *Nature* **1970,** *227* (5258), 561-3.

4. Fleischmann, R. D.; Adams, M. D.; White, O.; Clayton, R. A.; Kirkness, E. F.; Kerlavage, A. R.; Bult, C. J.; Tomb, J. F.; Dougherty, B. A.; Merrick, J. M.; et al., Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **1995,** *269* (5223), 496-512.

5. Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; Gocayne, J. D.; Amanatides, P.; Ballew, R. M.; Huson, D. H.; Wortman, J. R.; Zhang, Q.; Kodira, C. D.; Zheng, X. H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P. D.; Zhang, J.; Gabor Miklos, G. L.; Nelson, C.; Broder, S.; Clark, A. G.; Nadeau, J.; McKusick, V. A.; Zinder, N.; Levine, A. J.; Roberts, R. J.; Simon, M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florea, L.; Halpern, A.; Hannenhalli, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z.; Di Francesco, V.; Dunn, P.; Eilbeck, K.; Evangelista, C.; Gabrielian, A. E.; Gan, W.; Ge, W.; Gong, F.; Gu, Z.; Guan, P.; Heiman, T. J.; Higgins, M. E.; Ji, R. R.; Ke, Z.; Ketchum, K. A.; Lai, Z.; Lei, Y.; Li, Z.; Li, J.; Liang, Y.; Lin, X.; Lu, F.; Merkulov, G. V.; Milshina, N.; Moore, H. M.; Naik, A. K.; Narayan, V. A.; Neelam, B.; Nusskern, D.; Rusch, D. B.; Salzberg, S.; Shao, W.; Shue, B.; Sun, J.; Wang, Z.; Wang, A.; Wang, X.; Wang, J.; Wei, M.; Wides, R.; Xiao, C.; Yan, C.; Yao, A.; Ye, J.; Zhan, M.; Zhang, W.; Zhang, H.; Zhao, Q.; Zheng, L.; Zhong, F.; Zhong, W.; Zhu, S.; Zhao, S.; Gilbert, D.; Baumhueter, S.; Spier, G.; Carter, C.; Cravchik, A.; Woodage, T.; Ali, F.; An, H.; Awe, A.; Baldwin, D.; Baden, H.; Barnstead, M.; Barrow, I.; Beeson, K.; Busam, D.; Carver, A.; Center, A.; Cheng, M. L.; Curry, L.; Danaher, S.; Davenport, L.; Desilets, R.; Dietz, S.; Dodson, K.; Doup, L.; Ferriera, S.; Garg, N.; Gluecksmann, A.; Hart, B.; Haynes, J.; Haynes, C.; Heiner, C.; Hladun, S.; Hostin, D.; Houck, J.; Howland, T.; Ibegwam, C.; Johnson, J.; Kalush, F.; Kline, L.; Koduru, S.; Love, A.; Mann, F.; May, D.; McCawley, S.; McIntosh, T.; McMullen, I.; Moy, M.; Moy, L.; Murphy, B.; Nelson, K.; Pfannkoch, C.; Pratts, E.; Puri, V.; Qureshi, H.; Reardon, M.; Rodriguez, R.; Rogers, Y. H.; Romblad, D.; Ruhfel, B.; Scott, R.; Sitter, C.; Smallwood, M.; Stewart, E.; Strong, R.; Suh, E.; Thomas, R.; Tint, N. N.; Tse, S.; Vech, C.; Wang, G.; Wetter, J.; Williams, S.; Williams, M.; Windsor, S.; Winn-Deen, E.; Wolfe, K.; Zaveri, J.; Zaveri, K.; Abril, J. F.; Guigo, R.; Campbell, M. J.; Sjolander, K. V.; Karlak, B.; Kejariwal, A.; Mi, H.; Lazareva, B.; Hatton, T.; Narechania, A.; Diemer, K.; Muruganujan, A.; Guo, N.; Sato,

S.; Bafna, V.; Istrail, S.; Lippert, R.; Schwartz, R.; Walenz, B.; Yooseph, S.; Allen, D.; Basu, A.; Baxendale, J.; Blick, L.; Caminha, M.; Carnes-Stine, J.; Caulk, P.; Chiang, Y. H.; Coyne, M.; Dahlke, C.; Mays, A.; Dombroski, M.; Donnelly, M.; Ely, D.; Esparham, S.; Fosler, C.; Gire, H.; Glanowski, S.; Glasser, K.; Glodek, A.; Gorokhov, M.; Graham, K.; Gropman, B.; Harris, M.; Heil, J.; Henderson, S.; Hoover, J.; Jennings, D.; Jordan, C.; Jordan, J.; Kasha, J.; Kagan, L.; Kraft, C.; Levitsky, A.; Lewis, M.; Liu, X.; Lopez, J.; Ma, D.; Majoros, W.; McDaniel, J.; Murphy, S.; Newman, M.; Nguyen, T.; Nguyen, N.; Nodell, M.; Pan, S.; Peck, J.; Peterson, M.; Rowe, W.; Sanders, R.; Scott, J.; Simpson, M.; Smith, T.; Sprague, A.; Stockwell, T.; Turner, R.; Venter, E.; Wang, M.; Wen, M.; Wu, D.; Wu, M.; Xia, A.; Zandieh, A.; Zhu, X., The sequence of the human genome. *Science* **2001,** *291* (5507), 1304-51.

6. Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J. P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J. C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R. H.; Wilson, R. K.; Hillier, L. W.; McPherson, J. D.; Marra, M. A.; Mardis, E. R.; Fulton, L. A.; Chinwalla, A. T.; Pepin, K. H.; Gish, W. R.; Chissoe, S. L.; Wendl, M. C.; Delehaunty, K. D.; Miner, T. L.; Delehaunty, A.; Kramer, J. B.; Cook, L. L.; Fulton, R. S.; Johnson, D. L.; Minx, P. J.; Clifton, S. W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J. F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R. A.; Muzny, D. M.; Scherer, S. E.; Bouck, J. B.; Sodergren, E. J.; Worley, K. C.; Rives, C. M.; Gorrell, J. H.; Metzker, M. L.; Naylor, S. L.; Kucherlapati, R. S.; Nelson, D. L.; Weinstock, G. M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Smith, D. R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H. M.; Dubois, J.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R. W.; Federspiel, N. A.; Abola, A. P.; Proctor, M. J.; Myers, R. M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D. R.; Olson, M. V.; Kaul, R.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G. A.; Athanasiou, M.; Schultz, R.; Roe, B. A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W. R.; de la Bastide, M.; Dedhia, N.; Blocker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J. A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D. G.; Burge, C. B.; Cerutti, L.; Chen, H. C.; Church, D.; Clamp, M.; Copley, R. R.; Doerks, T.; Eddy, S. R.; Eichler, E. E.; Furey, T. S.; Galagan, J.; Gilbert, J. G.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L. S.; Jones, T. A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent,

W. J.; Kitts, P.; Koonin, E. V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T. M.; McLysaght, A.; Mikkelsen, T.; Moran, J. V.; Mulder, N.; Pollara, V. J.; Ponting, C. P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A. F.; Stupka, E.; Szustakowski, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y. I.; Wolfe, K. H.; Yang, S. P.; Yeh, R. F.; Collins, F.; Guyer, M. S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K. A.; Patrinos, A.; Morgan, M. J.; de Jong, P.; Catanese, J. J.; Osoegawa, K.; Shizuya, H.; Choi, S.; Chen, Y. J., Initial sequencing and analysis of the human genome. *Nature* **2001,** *409* (6822), 860-921.

7. Lander, E. S.; Weinberg, R. A., Genomics: journey to the center of biology. *Science* **2000,** *287* (5459), 1777-82.

8. Pagani, I.; Liolios, K.; Jansson, J.; Chen, I. M.; Smirnova, T.; Nosrat, B.; Markowitz, V. M.; Kyrpides, N. C., The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **2012,** *40* (Database issue), D571-9.

9. Vukmirovic, O. G.; Tilghman, S. M., Exploring genome space. *Nature* **2000,** *405* (6788), 820-2.

10. Baker, M., Functional genomics: The changes that count. *Nature* **2012,** *482* (7384), 257, 259-62.

11. David, L.; Huber, W.; Granovskaia, M.; Toedling, J.; Palm, C. J.; Bofkin, L.; Jones, T.; Davis, R. W.; Steinmetz, L. M., A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* **2006,** *103* (14), 5320-5.

12. Yamada, K.; Lim, J.; Dale, J. M.; Chen, H.; Shinn, P.; Palm, C. J.; Southwick, A. M.; Wu, H. C.; Kim, C.; Nguyen, M.; Pham, P.; Cheuk, R.; Karlin-Newmann, G.; Liu, S. X.; Lam, B.; Sakano, H.; Wu, T.; Yu, G.; Miranda, M.; Quach, H. L.; Tripp, M.; Chang, C. H.; Lee, J. M.; Toriumi, M.; Chan, M. M.; Tang, C. C.; Onodera, C. S.; Deng, J. M.; Akiyama, K.; Ansari, Y.; Arakawa, T.; Banh, J.; Banno, F.; Bowser, L.; Brooks, S.; Carninci, P.; Chao, Q.; Choy, N.; Enju, A.; Goldsmith, A. D.; Gurjal, M.; Hansen, N. F.; Hayashizaki, Y.; Johnson-Hopson, C.; Hsuan, V. W.; Iida, K.; Karnes, M.; Khan, S.; Koesema, E.; Ishida, J.; Jiang, P. X.; Jones, T.; Kawai, J.; Kamiya, A.; Meyers, C.; Nakajima, M.; Narusaka, M.; Seki, M.; Sakurai, T.; Satou, M.; Tamse, R.; Vaysberg, M.; Wallender, E. K.; Wong, C.; Yamamura, Y.; Yuan, S.; Shinozaki, K.; Davis, R. W.; Theologis, A.; Ecker, J. R., Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* **2003,** *302* (5646), 842-6.

13. Bertone, P.; Stolc, V.; Royce, T. E.; Rozowsky, J. S.; Urban, A. E.; Zhu, X.; Rinn, J. L.; Tongprasit, W.; Samanta, M.; Weissman, S.; Gerstein, M.; Snyder, M., Global

identification of human transcribed sequences with genome tiling arrays. *Science* **2004,** *306* (5705), 2242-6.

14. Nookaew, I.; Papini, M.; Pornputtapong, N.; Scalcinati, G.; Fagerberg, L.; Uhlen, M.; Nielsen, J., A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Res* **2012,** *40* (20), 10084-97.

15. Wang, Z.; Gerstein, M.; Snyder, M., RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **2009,** *10* (1), 57-63.

16. Wilkins, M. R.; Pasquali, C.; Appel, R. D.; Ou, K.; Golaz, O.; Sanchez, J. C.; Yan, J. X.; Gooley, A. A.; Hughes, G.; Humphery-Smith, I.; Williams, K. L.; Hochstrasser, D. F., From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N Y)* **1996,** *14* (1), 61-5.

17. O'Farrell, P. H., High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **1975,** *250* (10), 4007-21.

18. Klose, J., Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* **1975,** *26* (3), 231-43.

19. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M., Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989,** *246* (4926), 64-71.

20. Karas, M.; Hillenkamp, F., Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* **1988,** *60* (20), 2299-301.

21. Wollnik, H., Time-of-Flight Mass Analyzers. *Mass Spectrometry Reviews* **1993,** *12* (2), 89-114.

22. March, R. E., Quadrupole ion trap mass spectrometry: a view at the turn of the century. *International Journal of Mass Spectrometry* **2000,** *200* (1-3), 285-312.

23. Schwartz, J. C.; Jardine, I., Quadrupole ion trap mass spectrometry. *Methods Enzymol* **1996,** *270*, 552-86.

24. Spengler, B.; Kirsch, D.; Kaufmann, R.; Lemoine, J., Structure-Analysis of Branched Oligosaccharides Using Post-Source Decay in Matrix-Assisted Laser-Desorption Ionization Mass-Spectrometry. *Org Mass Spectrom* **1994,** *29* (12), 782-787.

25. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. *Nature* **2003,** *422* (6928), 198-207.

26. Hunt, D. F.; Henderson, R. A.; Shabanowitz, J.; Sakaguchi, K.; Michel, H.; Sevilir, N.; Cox, A. L.; Appella, E.; Engelhard, V. H., Characterization of Peptides Bound to the Class-I Mhc Molecule Hla-A2.1 by Mass-Spectrometry. *Science* **1992,** *255* (5049), 1261-1263.

27. Tipton, J. D.; Tran, J. C.; Catherman, A. D.; Ahlf, D. R.; Durbin, K. R.; Kelleher, N. L., Analysis of Intact Protein Isoforms by Mass Spectrometry. *Journal of Biological Chemistry* **2011,** *286* (29), 25451-25458.

28. Loo, J. A.; Edmonds, C. G.; Smith, R. D., Primary Sequence Information from Intact Proteins by Electrospray Ionization Tandem Mass-Spectrometry. *Science* **1990,** *248* (4952), 201-204.

29. Loo, J. A.; Quinn, J. P.; Ryu, S. I.; Henry, K. D.; Senko, M. W.; Mclafferty, F. W., High-Resolution Tandem Mass-Spectrometry of Large Biomolecules. *P Natl Acad Sci USA* **1992,** *89* (1), 286-289.

30. Senko, M. W.; Beu, S. C.; Mclafferty, F. W., High-Resolution Tandem Mass-Spectrometry of Carbonic-Anhydrase. *Analytical Chemistry* **1994,** *66* (3), 415-417.

31. Syka, J. E.; Marto, J. A.; Bai, D. L.; Horning, S.; Senko, M. W.; Schwartz, J. C.; Ueberheide, B.; Garcia, B.; Busby, S.; Muratore, T.; Shabanowitz, J.; Hunt, D. F., Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J Proteome Res* **2004,** *3* (3), 621-6.

32. Parks, B. A.; Jiang, L.; Thomas, P. M.; Wenger, C. D.; Roth, M. J.; Boyne, M. T.; Burke, P. V.; Kwast, K. E.; Kelleher, N. L., Top-down proteomics on a chromatographic time scale using linear ion trap Fourier transform hybrid mass spectrometers. *Analytical Chemistry* **2007,** *79* (21), 7984-7991.

33. Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L., Mapping

intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011,** *480* (7376), 254-8.

34. Bunger, M. K.; Cargile, B. J.; Ngunjiri, A.; Bundy, J. L.; Stephenson, J. L., Jr., Automated proteomics of E. coli via top-down electron-transfer dissociation mass spectrometry. *Anal Chem* **2008,** *80* (5), 1459-67.

35. Kellie, J. F.; Tran, J. C.; Lee, J. E.; Ahlf, D. R.; Thomas, H. M.; Ntai, I.; Catherman, A. D.; Durbin, K. R.; Zamdborg, L.; Vellaichamy, A.; Thomas, P. M.; Kelleher, N. L., The emerging process of Top Down mass spectrometry for protein analysis: biomarkers, protein-therapeutics, and achieving high throughput. *Mol Biosyst* **2010,** *6* (9), 1532-9.

36. Compton, P. D.; Zamdborg, L.; Thomas, P. M.; Kelleher, N. L., On the scalability and requirements of whole protein mass spectrometry. *Anal Chem* **2011,** *83* (17), 6868-74.

37. Kelleher, N. L.; Lin, H. Y.; Valaskovic, G. A.; Aaserud, D. J.; Fridriksson, E. K.; McLafferty, F. W., Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry. *Journal of the American Chemical Society* **1999,** *121* (4), 806-812.

38. Yates, J. R., Mass spectral analysis in proteomics. *Annu Rev Bioph Biom* **2004,** *33*, 297-316.

39. Henzel, W. J.; Watanabe, C.; Stults, J. T., Protein identification: The origins of peptide mass fingerprinting. *J Am Soc Mass Spectr* **2003,** *14* (9), 931-942.

40. Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C., Identification of 2-D Gel Proteins at the Femtomole Level by Molecular-Mass Searching of Peptide-Fragments in a Protein-Sequence Database. *Techniques in Protein Chemistry V* **1994**, 3-9.

41. Mann, M.; Wilm, M., Error Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Analytical Chemistry* **1994,** *66* (24), 4390-4399.

42. Eng, J. K.; Mccormack, A. L.; Yates, J. R., An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. *J Am Soc Mass Spectr* **1994,** *5* (11), 976-989.

43. Hunt, D. F.; Buko, A. M.; Ballard, J. M.; Shabanowitz, J.; Giordani, A. B., Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer. *Biomedical mass spectrometry* **1981,** *8* (9), 397-408.

44. Shukla, A. K.; Futrell, J. H., Tandem mass spectrometry: dissociation of ions by collisional activation. *J Mass Spectrom* **2000,** *35* (9), 1069-90.

45. Jonscher, K. R.; Yates, J. R., 3rd, The quadrupole ion trap mass spectrometer--a small solution to a big challenge. *Anal Biochem* **1997,** *244* (1), 1-15.

46. Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A., Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom* **2000,** *35* (12), 1399-406.

47. Boyd, R.; Somogyi, A., The mobile proton hypothesis in fragmentation of protonated peptides: a perspective. *J Am Soc Mass Spectrom* **2010,** *21* (8), 1275-8.

48. Biemann, K., Nomenclature for Peptide Fragment Ions (Positive-Ions). *Method Enzymol* **1990,** *193*, 886-887.

49. Nau, H.; Kelley, J. A.; Biemann, K., Determination of the amino acid sequence of the C-terminal cyanogen bromide fragment of actin by computer-assisted gas chromatography--mass spectrometry. *J Am Chem Soc* **1973,** *95* (21), 7162-4.

50. Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* **2003,** *17* (20), 2337-42.

51. Mo, L.; Dutta, D.; Wan, Y.; Chen, T., MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. *Anal Chem* **2007,** *79* (13), 4870-8.

52. Frank, A.; Pevzner, P., PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* **2005,** *77* (4), 964-73.

53. Nesvizhskii, A. I.; Roos, F. F.; Grossmann, J.; Vogelzang, M.; Eddes, J. S.; Gruissem, W.; Baginsky, S.; Aebersold, R., Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics* **2006,** *5* (4), 652-70.

54. Tabb, D. L.; Saraf, A.; Yates, J. R., 3rd, GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem* **2003,** *75* (23), 6415-21.

55. Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V., InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* **2005,** *77* (14), 4626-39.

56. Hernandez, P.; Gras, R.; Frey, J.; Appel, R. D., Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics* **2003,** *3* (6), 870-8.

57. Tabb, D. L.; Ma, Z. Q.; Martin, D. B.; Ham, A. J.; Chambers, M. C., DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J Proteome Res* **2008,** *7* (9), 3838-46.

58. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999,** *20* (18), 3551-67.

59. Clauser, K. R.; Baker, P.; Burlingame, A. L., Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem* **1999,** *71* (14), 2871-82.

60. Zhang, N.; Aebersold, R.; Schwikowski, B., ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2002,** *2* (10), 1406-12.

61. Craig, R.; Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004,** *20* (9), 1466-7.

62. Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J., OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* **2003,** *3* (8), 1454-63.

63. Tabb, D. L.; Fernando, C. G.; Chambers, M. C., MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* **2007,** *6* (2), 654-61.

64. Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H., Open mass spectrometry search algorithm. *J Proteome Res* **2004,** *3* (5), 958-64.

65. Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R., Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007,** *7* (5), 655-67.

66. Balgley, B. M.; Laudeman, T.; Yang, L.; Song, T.; Lee, C. S., Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. *Mol Cell Proteomics* **2007,** *6* (9), 1599-608.

67. Benjamini, Y.; Drai, D.; Elmer, G.; Kafkafi, N.; Golani, I., Controlling the false discovery rate in behavior genetics research. *Behavioural brain research* **2001,** *125* (1-2), 279-84.

68. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007,** *4* (3), 207-14.

69. Resing, K. A.; Meyer-Arendt, K.; Mendoza, A. M.; Aveline-Wolf, L. D.; Jonscher, K. R.; Pierce, K. G.; Old, W. M.; Cheung, H. T.; Russell, S.; Wattawa, J. L.; Goehle, G. R.; Knight, R. D.; Ahn, N. G., Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal Chem* **2004,** *76* (13), 3556-68.

70. Weatherly, D. B.; Atwood, J. A., 3rd; Minning, T. A.; Cavola, C.; Tarleton, R. L.; Orlando, R., A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. *Mol Cell Proteomics* **2005,** *4* (6), 762-72.

71. Yang, X.; Dondeti, V.; Dezube, R.; Maynard, D. M.; Geer, L. Y.; Epstein, J.; Chen, X.; Markey, S. P.; Kowalak, J. A., DBParser: web-based software for shotgun proteomic data analyses. *J Proteome Res* **2004,** *3* (5), 1002-8.

72. Rappsilber, J.; Mann, M., What does it mean to identify a protein in proteomics? *Trends Biochem Sci* **2002,** *27* (2), 74-8.

73. Nesvizhskii, A. I.; Aebersold, R., Interpretation of shotgun proteomic data - The protein inference problem. *Molecular & Cellular Proteomics* **2005,** *4* (10), 1419-1440.

74. Omenn, G. S.; States, D. J.; Adamski, M.; Blackwell, T. W.; Menon, R.; Hermjakob, H.; Apweiler, R.; Haab, B. B.; Simpson, R. J.; Eddes, J. S.; Kapp, E. A.; Moritz, R. L.; Chan, D. W.; Rai, A. J.; Admon, A.; Aebersold, R.; Eng, J.; Hancock, W. S.; Hefta, S. A.; Meyer, H.; Paik, Y. K.; Yoo, J. S.; Ping, P.; Pounds, J.; Adkins, J.; Qian, X.; Wang, R.; Wasinger, V.; Wu, C. Y.; Zhao, X.; Zeng, R.; Archakov, A.; Tsugita, A.; Beer, I.; Pandey, A.; Pisano, M.; Andrews, P.; Tammen, H.; Speicher, D. W.; Hanash, S.

M., Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **2005,** *5* (13), 3226-45.

75. Jaffe, J. D.; Stange-Thomann, N.; Smith, C.; DeCaprio, D.; Fisher, S.; Butler, J.; Calvo, S.; Elkins, T.; FitzGerald, M. G.; Hafez, N.; Kodira, C. D.; Major, J.; Wang, S.; Wilkinson, J.; Nicol, R.; Nusbaum, C.; Birren, B.; Berg, H. C.; Church, G. M., The complete genome and proteome of Mycoplasma mobile. *Genome Res* **2004,** *14* (8), 1447-61.

76. Jaffe, J. D.; Berg, H. C.; Church, G. M., Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **2004,** *4* (1), 59-77.

77. Schrimpf, S. P.; Weiss, M.; Reiter, L.; Ahrens, C. H.; Jovanovic, M.; Malmstrom, J.; Brunner, E.; Mohanty, S.; Lercher, M. J.; Hunziker, P. E.; Aebersold, R.; von Mering, C.; Hengartner, M. O., Comparative Functional Analysis of the Caenorhabditis elegans and Drosophila melanogaster Proteomes. *Plos Biology* **2009,** *7* (3), 616-627.

78. Picotti, P.; Clement-Ziza, M.; Lam, H.; Campbell, D. S.; Schmidt, A.; Deutsch, E. W.; Rost, H.; Sun, Z.; Rinner, O.; Reiter, L.; Shen, Q.; Michaelson, J. J.; Frei, A.; Alberti, S.; Kusebauch, U.; Wollscheid, B.; Moritz, R. L.; Beyer, A.; Aebersold, R., A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* **2013,** *494* (7436), 266-70.

79. Baerenfaller, K.; Grossmann, J.; Grobei, M. A.; Hull, R.; Hirsch-Hoffmann, M.; Yalovsky, S.; Zimmermann, P.; Grossniklaus, U.; Gruissem, W.; Baginsky, S., Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* **2008,** *320* (5878), 938-41.

80. Ong, S. E.; Mann, M., Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* **2005,** *1* (5), 252-62.

81. Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* **2007,** *389* (4), 1017-31.

82. Strittmatter, E. F.; Ferguson, P. L.; Tang, K.; Smith, R. D., Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry. *J Am Soc Mass Spectrom* **2003,** *14* (9), 980-91.

83. Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G. Z.; McKenna, T.; Nold, M. J.; Richardson, K.; Young, P.; Geromanos, S., Quantitative

proteomic analysis by accurate mass retention time pairs. *Anal Chem* **2005,** *77* (7), 2187-200.

84. Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **2001,** *19* (3), 242-7.

85. Gilchrist, A.; Au, C. E.; Hiding, J.; Bell, A. W.; Fernandez-Rodriguez, J.; Lesimple, S.; Nagaya, H.; Roy, L.; Gosline, S. J.; Hallett, M.; Paiement, J.; Kearney, R. E.; Nilsson, T.; Bergeron, J. J., Quantitative proteomics analysis of the secretory pathway. *Cell* **2006,** *127* (6), 1265-81.

86. Choi, H.; Fermin, D.; Nesvizhskii, A. I., Significance analysis of spectral count data in label-free shotgun proteomics. *Mol Cell Proteomics* **2008,** *7* (12), 2373-85.

87. Lahm, H. W.; Langen, H., Mass spectrometry: a tool for the identification of proteins separated by gels. *Electrophoresis* **2000,** *21* (11), 2105-14.

88. Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M., Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **2002,** *1* (5), 376-86.

89. Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R., Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **1999,** *17* (10), 994-9.

90. Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **2004,** *3* (12), 1154-69.

91. Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C., Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **2003,** *75* (8), 1895-904.

92. Dephoure, N.; Gygi, S. P., Hyperplexing: a method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin in yeast. *Sci Signal* **2012,** *5* (217), rs2.

93. Domon, B.; Aebersold, R., Options and considerations when selecting a quantitative proteomics strategy. *Nat Biotechnol* **2010,** *28* (7), 710-21.

94. Gerber, S. A.; Rush, J.; Stemman, O.; Kirschner, M. W.; Gygi, S. P., Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* **2003,** *100* (12), 6940-5.

95. Rivers, J.; Simpson, D. M.; Robertson, D. H.; Gaskell, S. J.; Beynon, R. J., Absolute multiplexed quantitative analysis of protein expression during muscle development using QconCAT. *Mol Cell Proteomics* **2007,** *6* (8), 1416-27.

96. Picotti, P.; Aebersold, R., Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Methods* **2012,** *9* (6), 555-66.

97. Lange, V.; Picotti, P.; Domon, B.; Aebersold, R., Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol* **2008,** *4*, 222.

98. Picotti, P.; Bodenmiller, B.; Mueller, L. N.; Domon, B.; Aebersold, R., Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **2009,** *138* (4), 795-806.

99. Picotti, P.; Bodenmiller, B.; Aebersold, R., Proteomics meets the scientific method. *Nat Methods* **2013,** *10* (1), 24-7.

100. Chum, H. L.; Overend, R. P., Biomass and renewable fuels. *Fuel Process Technol* **2001,** *71* (1-3), 187-195.

101. Faaij, A. P. C., Bio-energy in Europe: changing technology choices. *Energ Policy* **2006,** *34* (3), 322-342.

102. Himmel, M. E.; Ding, S. Y.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D., Biomass recalcitrance: Engineering plants and enzymes for biofuels production. *Science* **2007,** *315* (5813), 804-807.

103. Hammerschlag, R., Ethanol's energy return on investment: A survey of the literature 1990 - Present. *Environmental Science & Technology* **2006,** *40* (6), 1744-1750.

104. Chang, M. C., Harnessing energy from plant biomass. *Curr Opin Chem Biol* **2007,** *11* (6), 677-84.

105. Kumar, R.; Singh, S.; Singh, O. V., Bioconversion of lignocellulosic biomass: biochemical and molecular perspectives. *J Ind Microbiol Biotechnol* **2008,** *35* (5), 377-91.

106. Ragauskas, A. J.; Williams, C. K.; Davison, B. H.; Britovsek, G.; Cairney, J.; Eckert, C. A.; Frederick, W. J., Jr.; Hallett, J. P.; Leak, D. J.; Liotta, C. L.; Mielenz, J. R.; Murphy, R.; Templer, R.; Tschaplinski, T., The path forward for biofuels and biomaterials. *Science* **2006,** *311* (5760), 484-9.

107. Beckham, G. T.; Matthews, J. F.; Peters, B.; Bomble, Y. J.; Himmel, M. E.; Crowley, M. F., Molecular-level origins of biomass recalcitrance: decrystallization free energies for four common cellulose polymorphs. *J Phys Chem B* **2011,** *115* (14), 4118-27.

108. Himmel, M. E.; Ding, S. Y.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D., Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* **2007,** *315* (5813), 804-7.

109. Lynd, L. R.; Elander, R. T.; Wyman, C. E., Likely features and costs of mature biomass ethanol technology. *Appl Biochem Biotech* **1996,** *57-8*, 741-761.

110. Ye, X.; Busov, V.; Zhao, N.; Meilan, R.; McDonnell, L. M.; Coleman, H. D.; Mansfield, S. D.; Chen, F.; Li, Y.; Cheng, Z. M., Transgenic Populus Trees for Forest Products, Bioenergy, and Functional Genomics. *Crit Rev Plant Sci* **2011,** *30* (5), 415-434.

111. Tuskan, G. A., Short-rotation woody crop supply systems in the United States: What do we know and what do we need to know? *Biomass Bioenerg* **1998,** *14* (4), 307-315.

112. Tuskan, G. A.; DiFazio, S.; Jansson, S.; Bohlmann, J.; Grigoriev, I.; Hellsten, U.; Putnam, N.; Ralph, S.; Rombauts, S.; Salamov, A.; Schein, J.; Sterck, L.; Aerts, A.; Bhalerao, R. R.; Bhalerao, R. P.; Blaudez, D.; Boerjan, W.; Brun, A.; Brunner, A.; Busov, V.; Campbell, M.; Carlson, J.; Chalot, M.; Chapman, J.; Chen, G. L.; Cooper, D.; Coutinho, P. M.; Couturier, J.; Covert, S.; Cronk, Q.; Cunningham, R.; Davis, J.; Degroeve, S.; Dejardin, A.; Depamphilis, C.; Detter, J.; Dirks, B.; Dubchak, I.; Duplessis, S.; Ehlting, J.; Ellis, B.; Gendler, K.; Goodstein, D.; Gribskov, M.; Grimwood, J.; Groover, A.; Gunter, L.; Hamberger, B.; Heinze, B.; Helariutta, Y.; Henrissat, B.; Holligan, D.; Holt, R.; Huang, W.; Islam-Faridi, N.; Jones, S.; Jones-Rhoades, M.; Jorgensen, R.; Joshi, C.; Kangasjarvi, J.; Karlsson, J.; Kelleher, C.; Kirkpatrick, R.; Kirst, M.; Kohler, A.; Kalluri, U.; Larimer, F.; Leebens-Mack, J.; Leple, J. C.; Locascio, P.; Lou, Y.; Lucas, S.; Martin, F.; Montanini, B.; Napoli, C.; Nelson, D. R.; Nelson, C.; Nieminen, K.; Nilsson, O.; Pereda, V.; Peter, G.; Philippe, R.; Pilate, G.; Poliakov, A.; Razumovskaya, J.; Richardson, P.; Rinaldi, C.; Ritland, K.; Rouze, P.; Ryaboy, D.;

Schmutz, J.; Schrader, J.; Segerman, B.; Shin, H.; Siddiqui, A.; Sterky, F.; Terry, A.; Tsai, C. J.; Uberbacher, E.; Unneberg, P.; Vahala, J.; Wall, K.; Wessler, S.; Yang, G.; Yin, T.; Douglas, C.; Marra, M.; Sandberg, G.; Van de Peer, Y.; Rokhsar, D., The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **2006,** *313* (5793), 1596-1604.

113. Wullschleger, S. D.; Jansson, S.; Taylor, G., Genomics and forest biology: Populus emerges as the perennial favorite. *Plant Cell* **2002,** *14* (11), 2651-2655.

114. Slavov, G. T.; DiFazio, S. P.; Martin, J.; Schackwitz, W.; Muchero, W.; Rodgers-Melnick, E.; Lipphardt, M. F.; Pennacchio, C. P.; Hellsten, U.; Pennacchio, L. A.; Gunter, L. E.; Ranjan, P.; Vining, K.; Pomraning, K. R.; Wilhelm, L. J.; Pellegrini, M.; Mockler, T. C.; Freitag, M.; Geraldes, A.; El-Kassaby, Y. A.; Mansfield, S. D.; Cronk, Q. C.; Douglas, C. J.; Strauss, S. H.; Rokhsar, D.; Tuskan, G. A., Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree Populus trichocarpa. *New Phytol* **2012,** *196* (3), 713-25.

115. Neale, D. B.; Kremer, A., Forest tree genomics: growing resources and applications. *Nature Reviews Genetics* **2011,** *12* (2), 111-122.

116. Porth, I.; Klapste, J.; Skyba, O.; Lai, B. S. K.; Geraldes, A.; Muchero, W.; Tuskan, G. A.; Douglas, C. J.; El-Kassaby, Y. A.; Mansfield, S. D., Populus trichocarpa cell wall chemistry and ultrastructure trait variation, genetic control and genetic correlations. *New Phytologist* **2013,** *197* (3), 777-790.

117. Wullschleger, S. D.; Weston, D. J.; Difazio, S. P.; Tuskan, G. A., Revisiting the sequencing of the first tree genome: Populus trichocarpa. *Tree Physiol* **2012**.

118. Geraldes, A.; Pang, J.; Thiessen, N.; Cezard, T.; Moore, R.; Zhao, Y.; Tam, A.; Wang, S.; Friedmann, M.; Birol, I.; Jones, S. J.; Cronk, Q. C.; Douglas, C. J., SNP discovery in black cottonwood (Populus trichocarpa) by population transcriptome resequencing. *Mol Ecol Resour* **2011,** *11 Suppl 1*, 81-92.

119. Hsu, C. Y.; Adams, J. P.; Kim, H.; No, K.; Ma, C.; Strauss, S. H.; Drnevich, J.; Vandervelde, L.; Ellis, J. D.; Rice, B. M.; Wickett, N.; Gunter, L. E.; Tuskan, G. A.; Brunner, A. M.; Page, G. P.; Barakat, A.; Carlson, J. E.; DePamphilis, C. W.; Luthe, D. S.; Yuceer, C., FLOWERING LOCUS T duplication coordinates reproductive and vegetative growth in perennial poplar. *Proc Natl Acad Sci U S A* **2011,** *108* (26), 10756-61.

120. Hsu, C. Y.; Liu, Y. X.; Luthe, D. S.; Yuceer, C., Poplar FT2 shortens the juvenile phase and promotes seasonal flowering. *Plant Cell* **2006,** *18* (8), 1846-1861.

121. Gorg, A.; Weiss, W.; Dunn, M. J., Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **2004,** *4* (12), 3665-85.

122. Schiltz, S.; Gallardo, K.; Huart, M.; Negroni, L.; Sommerer, N.; Burstin, J., Proteome reference maps of vegetative tissues in pea. An investigation of nitrogen mobilization from leaves during seed filling. *Plant Physiol* **2004,** *135* (4), 2241-60.

123. Giavalisco, P.; Kapitza, K.; Kolasa, A.; Buhtz, A.; Kehr, J., Towards the proteome of Brassica napus phloem sap. *Proteomics* **2006,** *6* (3), 896-909.

124. Kieffer, P.; Dommes, J.; Hoffmann, L.; Hausman, J. F.; Renaut, J., Quantitative changes in protein expression of cadmium-exposed poplar plants. *Proteomics* **2008,** *8* (12), 2514-2530.

125. Plomion, C.; Lalanne, C.; Claverol, S.; Meddour, H.; Kohler, A.; Bogeat-Triboulot, M. B.; Barre, A.; Le Provost, G.; Dumazet, H.; Jacob, D.; Bastien, C.; Dreyer, E.; de Daruvar, A.; Guehl, J. M.; Schmitter, J. M.; Martin, F.; Bonneu, M., Mapping the proteome of poplar and application to the discovery of drought-stress responsive proteins. *Proteomics* **2006,** *6* (24), 6509-6527.

126. Lambert, J. P.; Ethier, M.; Smith, J. C.; Figeys, D., Proteomics: from gel based to gel free. *Anal Chem* **2005,** *77* (12), 3771-87.

127. Koller, A.; Washburn, M. P.; Lange, B. M.; Andon, N. L.; Deciu, C.; Haynes, P. A.; Hays, L.; Schieltz, D.; Ulaszek, R.; Wei, J.; Wolters, D.; Yates, J. R., Proteomic survey of metabolic pathways in rice. *P Natl Acad Sci USA* **2002,** *99* (18), 11969-11974.

128. Kalluri, U. C.; Hurst, G. B.; Lankford, P. K.; Ranjan, P.; Pelletier, D. A., Shotgun proteome profile of Populus developing xylem. *Proteomics* **2009,** *9* (21), 4871-4880.

129. Gion, J. M.; Lalanne, C.; Le Provost, G.; Ferry-Dumazet, H.; Paiva, J.; Chaumeil, P.; Frigerio, J. M.; Brach, J.; Barre, A.; de Daruvar, A.; Claverol, S.; Bonneu, M.; Sommerer, N.; Negroni, L.; Plomion, C., The proteome of maritime pine wood forming tissue. *Proteomics* **2005,** *5* (14), 3731-51.

130. Smith, P. K.; Krohn, R. I.; Hermanson, G. T.; Mallia, A. K.; Gartner, F. H.; Provenzano, M. D.; Fujimoto, E. K.; Goeke, N. M.; Olson, B. J.; Klenk, D. C., Measurement of Protein Using Bicinchoninic Acid. *Analytical Biochemistry* **1985,** *150* (1), 76-85.

131. Wiechelman, K. J.; Braun, R. D.; Fitzpatrick, J. D., Investigation of the Bicinchoninic Acid Protein Assay - Identification of the Groups Responsible for Color Formation. *Analytical Biochemistry* **1988,** *175* (1), 231-237.

132. Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd, Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **1999,** *17* (7), 676-82.

133. Giddings, J. C., Concepts and Comparisons in Multidimensional Separation. *J High Res Chromatog* **1987,** *10* (5), 319-323.

134. Guiochon, G.; Marchetti, N.; Mriziq, K.; Shalliker, R. A., Implementations of two-dimensional liquid chromatography. *Journal of Chromatography A* **2008,** *1189* (1-2), 109-168.

135. McLuckey, S. A.; Wells, J. M., Mass analysis at the advent of the 21st century. *Chem Rev* **2001,** *101* (2), 571-606.

136. Taylor, G., Disintegration of water drops in an electric field. *Proceedings of the Royal Society of London A* **1964,** *280*, 383-397.

137. Fenn, J. B., Ion Formation from Charged Droplets - Roles of Geometry, Energy, and Time. *J Am Soc Mass Spectr* **1993,** *4* (7), 524-535.

138. Kebarle, P.; Peschke, M., On the mechanisms by which the charged droplets produced by electrospray lead to gas phase ions. *Analytica Chimica Acta* **2000,** *406* (1), 11-35.

139. Rayleigh, L., On the equilibrium of liquid conducting masses charged with electricity. *Philosophical Magazine* **1882**, 184-186.

140. Tang, L.; Kebarle, P., Effect of the Conductivity of the Electrosprayed Solution on the Electrospray Current - Factors Determining Analyte Sensitivity in Electrospray Mass-Spectrometry. *Analytical Chemistry* **1991,** *63* (23), 2709-2715.

141. Kebarle, P.; Verkerk, U. H., Electrospray: From Ions in Solution to Ions in the Gas Phase, What We Know Now. *Mass Spectrometry Reviews* **2009,** *28* (6), 898-917.

142. Tang, L.; Kebarle, P., Dependence of Ion Intensity in Electrospray Mass-Spectrometry on the Concentration of the Analytes in the Electrosprayed Solution. *Analytical Chemistry* **1993,** *65* (24), 3654-3668.

143. Karas, M.; Bahr, U.; Dulcks, T., Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine. *Fresen J Anal Chem* **2000,** *366* (6-7), 669-676.

144. Wilm, M.; Mann, M., Analytical properties of the nanoelectrospray ion source. *Analytical Chemistry* **1996,** *68* (1), 1-8.

145. Valaskovic, G. A.; Kelleher, N. L.; McLafferty, F. W., Attomole protein characterization by capillary electrophoresis mass spectrometry. *Science* **1996,** *273* (5279), 1199-1202.

146. Schwartz, J. C.; Jardine, I., Quadrupole ion trap mass spectrometry. *High Resolution Separation and Analysis of Biological Macromolecules, Pt A* **1996,** *270*, 552-586.

147. Todd, J. F. J.; March, R. E., A retrospective review of the development and application of the quadrupole ion trap prior to the appearance of commercial instruments. *International Journal of Mass Spectrometry* **1999,** *191*, 9-35.

148. Schwartz, J. C.; Senko, M. W.; Syka, J. E. P., A two-dimensional quadrupole ion trap mass spectrometer. *J Am Soc Mass Spectr* **2002,** *13* (6), 659-669.

149. Olsen, J. V.; Schwartz, J. C.; Griep-Raming, J.; Nielsen, M. L.; Damoc, E.; Denisov, E.; Lange, O.; Remes, P.; Taylor, D.; Splendore, M.; Wouters, E. R.; Senko, M.; Makarov, A.; Mann, M.; Horning, S., A Dual Pressure Linear Ion Trap Orbitrap Instrument with Very High Sequencing Speed. *Molecular & Cellular Proteomics* **2009,** *8* (12), 2759-2769.

150. Second, T. P.; Blethrow, J. D.; Schwartz, J. C.; Merrihew, G. E.; MacCoss, M. J.; Swaney, D. L.; Russell, J. D.; Coon, J. J.; Zabrouskov, V., Dual-Pressure Linear Ion Trap Mass Spectrometer Improving the Analysis of Complex Protein Mixtures. *Analytical Chemistry* **2009,** *81* (18), 7757-7765.

151. McLachlan, D. W., Theory and Applications of Mathieu Functions. *Clarendon: Oxford* **1947**.

152. Stafford, G., Jr., Ion trap mass spectrometry: a personal perspective. *J Am Soc Mass Spectrom* **2002,** *13* (6), 589-96.

153. Park, C. H., Further study of electron multiplication in conventional continuous dynode electron multiplier. *Ieee Nucl Sci Conf R* **2004**, 1398-1400.

154. Hu, Q.; Noll, R. J.; Li, H.; Makarov, A.; Hardman, M.; Graham Cooks, R., The Orbitrap: a new mass spectrometer. *J Mass Spectrom* **2005,** *40* (4), 430-43.

155. Makarov, A., Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry* **2000,** *72* (6), 1156-1162.

156. Michalski, A.; Damoc, E.; Lange, O.; Denisov, E.; Nolting, D.; Muller, M.; Viner, R.; Schwartz, J.; Remes, P.; Belford, M.; Dunyach, J. J.; Cox, J.; Horning, S.; Mann, M.; Makarov, A., Ultra High Resolution Linear Ion Trap Orbitrap Mass Spectrometer (Orbitrap Elite) Facilitates Top Down LC MS/MS and Versatile Peptide Fragmentation Modes. *Molecular & Cellular Proteomics* **2012,** *11* (3).

157. Michalski, A.; Cox, J.; Mann, M., More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC-MS/MS. *Journal of Proteome Research* **2011,** *10* (4), 1785-1793.

158. Domon, B.; Aebersold, R., Mass spectrometry and protein analysis. *Science* **2006,** *312* (5771), 212-7.

159. Graumann, J.; Scheltema, R. A.; Zhang, Y.; Cox, J.; Mann, M., A Framework for Intelligent Data Acquisition and Real-Time Database Searching for Shotgun Proteomics. *Molecular & Cellular Proteomics* **2012,** *11* (3).

160. Beck, M.; Claassen, M.; Aebersold, R., Comprehensive proteomics. *Curr Opin Biotech* **2011,** *22* (1), 3-8.

161. Zhang, Y.; Wen, Z. H.; Washburn, M. P.; Florens, L., Effect of Dynamic Exclusion Duration on Spectral Count Based Quantitative Proteomics. *Analytical Chemistry* **2009,** *81* (15), 6317-6326.

162. Dasari, S.; Chambers, M. C.; Slebos, R. J.; Zimmerman, L. J.; Ham, A. J.; Tabb, D. L., TagRecon: high-throughput mutation identification through sequence tagging. *J Proteome Res* **2010,** *9* (4), 1716-26.

163. Tabb, D. L.; McDonald, W. H.; Yates, J. R., 3rd, DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* **2002,** *1* (1), 21-6.

164. Ma, Z. Q.; Dasari, S.; Chambers, M. C.; Litton, M. D.; Sobecki, S. M.; Zimmerman, L. J.; Halvey, P. J.; Schilling, B.; Drake, P. M.; Gibson, B. W.; Tabb, D. L., IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res* **2009,** *8* (8), 3872-81.

165. Wilmarth, P. A.; Tanner, S.; Dasari, S.; Nagalla, S. R.; Riviere, M. A.; Bafna, V.; Pevzner, P. A.; David, L. L., Age-related changes in human crystallins determined from comparative analysis of post-translational modifications in young and aged lens: does deamidation contribute to crystallin insolubility? *J Proteome Res* **2006,** *5* (10), 2554-66.

166. Edgar, R. C., Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010,** *26* (19), 2460-1.

167. Zybailov, B.; Coleman, M. K.; Florens, L.; Washburn, M. P., Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal Chem* **2005,** *77* (19), 6218-24.

168. Liu, H.; Sadygov, R. G.; Yates, J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **2004,** *76* (14), 4193-201.

169. Zybailov, B. L.; Florens, L.; Washburn, M. P., Quantitative shotgun proteomics using a protease with broad specificity and normalized spectral abundance factors. *Mol Biosyst* **2007,** *3* (5), 354-60.

170. Lochner, A.; Giannone, R. J.; Keller, M.; Antranikian, G.; Graham, D. E.; Hettich, R. L., Label-free quantitative proteomics for the extremely thermophilic bacterium Caldicellulosiruptor obsidiansis reveal distinct abundance patterns upon growth on cellobiose, crystalline cellulose, and switchgrass. *J Proteome Res* **2011,** *10* (12), 5302-14.

171. Deschamps, S.; Campbell, M. A., Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Mol Breeding* **2010,** *25* (4), 553-570.

172. Saito, K.; Matsuda, F., Metabolomics for Functional Genomics, Systems Biology, and Biotechnology. *Annual Review of Plant Biology, Vol 61* **2010,** *61*, 463-489.

173. Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R., Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology* **1999,** *19* (3), 1720-1730.

174. Fiehn, O.; Kopka, J.; Dormann, P.; Altmann, T.; Trethewey, R. N.; Willmitzer, L., Metabolite profiling for plant functional genomics. *Nature Biotechnology* **2000,** *18* (11), 1157-1161.

175. de Godoy, L. M. F.; Olsen, J. V.; Cox, J.; Nielsen, M. L.; Hubner, N. C.; Frohlich, F.; Walther, T. C.; Mann, M., Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **2008,** *455* (7217), 1251-U60.

176. Jansson, S.; Douglas, C. J., Populus: a model system for plant biology. *Annu Rev Plant Biol* **2007,** *58*, 435-58.

177. Bylesjo, M.; Nilsson, R.; Srivastava, V.; Gronlund, A.; Johansson, A. I.; Jansson, S.; Karlsson, J.; Moritz, T.; Wingsle, G.; Trygg, J., Integrated Analysis of Transcript, Protein and Metabolite Data To Study Lignin Biosynthesis in Hybrid Aspen. *Journal of Proteome Research* **2009,** *8* (1), 199-210.

178. Nilsson, R.; Bernfur, K.; Gustavsson, N.; Bygdell, J.; Wingsle, G.; Larsson, C., Proteomics of Plasma Membranes from Poplar Trees Reveals Tissue Distribution of Transporters, Receptors, and Proteins in Cell Wall Formation. *Molecular & Cellular Proteomics* **2010,** *9* (2), 368-387.

179. Visioli, G.; Marmiroli, M.; Marmiroli, N., Two-Dimensional Liquid Chromatography Technique Coupled with Mass Spectrometry Analysis to Compare the Proteomic Response to Cadmium Stress in Plants. *J Biomed Biotechnol* **2010**, -.

180. Nedelkov, D., Population proteomics: investigation of protein diversity in human populations. *Proteomics* **2008,** *8* (4), 779-86.

181. Zybailov, B.; Sun, Q.; van Wijk, K. J., Workflow for Large Scale Detection and Validation of Peptide Modifications by RPLC-LTQ-Orbitrap: Application to the Arabidopsis thaliana Leaf Proteome and an Online Modified Peptide Library. *Analytical Chemistry* **2009,** *81* (19), 8015-8024.

182. Foston, M.; Hubbell, C.; Samuel, R.; Jung, S.; Fan, H.; Ding, S.-Y.; Zeng, Y.; Jawdy, S.; Davis, M.; Sykes, S.; Gjersing, E.; Tuskan, G. A.; Kalluri, U.; Ragauskas, A. J., Chemical, ultrastructural and supramolecular analysis of tension wood in Populus tremula x alba as a model substrate for reduced recalcitrance. *Energy & Environmental Science Journal* **2011**.

183. Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **2003,** *75* (17), 4646-58.

184. Delalande, F.; Carapito, C.; Brizard, J. P.; Brugidou, C.; Van Dorsselaer, A., Multigenic families and proteomics: extended protein characterization as a tool for paralog gene identification. *Proteomics* **2005,** *5* (2), 450-60.

185. Black, D. L., Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* **2000,** *103* (3), 367-70.

186. Friso, G.; Majeran, W.; Huang, M.; Sun, Q.; van Wijk, K. J., Reconstruction of metabolic pathways, protein expression, and homeostasis machineries across maize bundle sheath and mesophyll chloroplasts: large-scale quantitative proteomics using the first maize genome assembly. *Plant Physiol* **2010,** *152* (3), 1219-50.

187. Meyer-Arendt, K.; Old, W. M.; Houel, S.; Renganathan, K.; Eichelberger, B.; Resing, K. A.; Ahn, N. G., IsoformResolver: A Peptide-Centric Algorithm for Protein Inference. *J Proteome Res* **2011**.

188. Edgar, R. C., Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010,** *26* (19), 2460-2461.

189. Wu, C. H.; Yeh, L. S.; Huang, H.; Arminski, L.; Castro-Alvear, J.; Chen, Y.; Hu, Z.; Kourtesis, P.; Ledley, R. S.; Suzek, B. E.; Vinayaka, C. R.; Zhang, J.; Barker, W. C., The Protein Information Resource. *Nucleic Acids Res* **2003,** *31* (1), 345-7.

190. Yang, X. H.; Tschaplinski, T. J.; Hurst, G. B.; Jawdy, S.; Abraham, P. E.; Lankford, P. K.; Adams, R. M.; Shah, M. B.; Hettich, R. L.; Lindquist, E.; Kalluri, U. C.; Gunter, L. E.; Pennacchio, C.; Tuskan, G. A., Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome Research* **2011,** *21* (4), 634-641.

191. Fisher, K.; Turner, S., PXY, a receptor-like kinase essential for maintaining polarity during plant vascular-tissue development. *Curr Biol* **2007,** *17* (12), 1061-6.

192. Hirakawa, Y.; Shinohara, H.; Kondo, Y.; Inoue, A.; Nakanomyo, I.; Ogawa, M.; Sawa, S.; Ohashi-Ito, K.; Matsubayashi, Y.; Fukuda, H., Non-cell-autonomous control of vascular stem cell fate by a CLE peptide/receptor system. *Proc Natl Acad Sci U S A* **2008,** *105* (39), 15208-13.

193. Yamamoto, R.; Fujioka, S.; Demura, T.; Takatsuto, S.; Yoshida, S.; Fukuda, H., Brassinosteroid levels increase drastically prior to morphogenesis of tracheary elements. *Plant Physiol* **2001,** *125* (2), 556-63.

194. Cano-Delgado, A.; Yin, Y. H.; Yu, C.; Vafeados, D.; Mora-Garcia, S.; Cheng, J. C.; Nam, K. H.; Li, J. M.; Chory, J., BRL1 and BRL3 are novel brassinosteroid receptors that function in vascular differentiation in Arabidopsis. *Development* **2004,** *131* (21), 5341-5351.

195. Nakamura, A.; Fujioka, S.; Sunohara, H.; Kamiya, N.; Hong, Z.; Inukai, Y.; Miura, K.; Takatsuto, S.; Yoshida, S.; Ueguchi-Tanaka, M.; Hasegawa, Y.; Kitano, H.; Matsuoka, M., The role of OsBRI1 and its homologous genes, OsBRL1 and OsBRL3, in rice. *Plant Physiology* **2006,** *140* (2), 580-590.

196. Fukuda, H., Signals that control plant vascular cell differentiation. *Nat Rev Mol Cell Bio* **2004,** *5* (5), 379-391.

197. Sauer, M.; Paciorek, T.; Benkova, E.; Friml, J., Immunocytochemical techniques for whole-mount in situ protein localization in plants. *Nat Protoc* **2006,** *1* (1), 98-103.

198. Scarpella, E.; Marcos, D.; Friml, J.; Berleth, T., Control of leaf vascular patterning by polar auxin transport. *Genes Dev* **2006,** *20* (8), 1015-27.

199. Tuskan, G. A.; Walsh, M. E., Short-rotation woody crop systems, atmospheric carbon dioxide and carbon management: A US case study. *Forest Chron* **2001,** *77* (2), 259-264.

200. Davison, B. H.; Drescher, S. R.; Tuskan, G. A.; Davis, M. F.; Nghiem, N. P., Variation of S/G ratio and lignin content in a Populus family influences the release of xylose by dilute acid hydrolysis. *Appl Biochem Biotech* **2006,** *130* (1-3), 427-435.

201. Dinus, R. J.; Payne, P.; Sewell, N. M.; Chiang, V. L.; Tuskan, G. A., Genetic modification of short rotation popular wood: Properties for ethanol fuel and fiber productions. *Crit Rev Plant Sci* **2001,** *20* (1), 51-69.

202. Sannigrahi, P.; Ragauskas, A. J.; Tuskan, G. A., Poplar as a feedstock for biofuels: A review of compositional characteristics. *Biofuel Bioprod Bior* **2010,** *4* (2), 209-226.

203. Cantarel, B. L.; Coutinho, P. M.; Rancurel, C.; Bernard, T.; Lombard, V.; Henrissat, B., The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Research* **2009,** *37*, D233-D238.

204. Aspeborg, H.; Schrader, J.; Coutinho, P. M.; Stam, M.; Kallas, A.; Djerbi, S.; Nilsson, P.; Denman, S.; Amini, B.; Sterky, F.; Master, E.; Sandberg, G.; Mellerowicz, E.; Sundberg, B.; Henrissat, B.; Teeri, T. T., Carbohydrate-active enzymes involved in the secondary cell wall biogenesis in hybrid aspen. *Plant Physiol* **2005,** *137* (3), 983-97.

205. Geisler-Lee, J.; Geisler, M.; Coutinho, P. M.; Segerman, B.; Nishikubo, N.; Takahashi, J.; Aspeborg, H.; Djerbi, S.; Master, E.; Andersson-Gunneras, S.; Sundberg, B.; Karpinski, S.; Teeri, T. T.; Kleczkowski, L. A.; Henrissat, B.; Mellerowicz, E. J.,

Poplar carbohydrate-active enzymes. Gene identification and expression analyses. *Plant Physiol* **2006,** *140* (3), 946-62.

206. Andersson-Gunneras, S.; Mellerowicz, E. J.; Love, J.; Segerman, B.; Ohmiya, Y.; Coutinho, P. M.; Nilsson, P.; Henrissat, B.; Moritz, T.; Sundberg, B., Biosynthesis of cellulose-enriched tension wood in Populus: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *Plant J* **2006,** *45* (2), 144-65.

207. Shi, R.; Sun, Y. H.; Li, Q. Z.; Heber, S.; Sederoff, R.; Chiang, V. L., Towards a Systems Approach for Lignin Biosynthesis in Populus trichocarpa: Transcript Abundance and Specificity of the Monolignol Biosynthetic Genes. *Plant and Cell Physiology* **2010,** *51* (1), 144-163.

208. Johnson, R. S.; Davis, M. T.; Taylor, J. A.; Patterson, S. D., Informatics for protein identification by mass spectrometry. *Methods* **2005,** *35* (3), 223-36.

209. Verberkmoes, N. C.; Hervey, W. J.; Shah, M.; Land, M.; Hauser, L.; Larimer, F. W.; Van Berkel, G. J.; Goeringer, D. E., Evaluation of "shotgun" proteomics for identification of biological threat agents in complex environmental matrixes: experimental simulations. *Anal Chem* **2005,** *77* (3), 923-32.

210. Bern, M.; Goldberg, D.; McDonald, W. H.; Yates, J. R., 3rd, Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics* **2004,** *20 Suppl 1*, i49-54.

211. Salmi, J.; Moulder, R.; Filen, J. J.; Nevalainen, O. S.; Nyman, T. A.; Lahesmaa, R.; Aittokallio, T., Quality classification of tandem mass spectrometry data. *Bioinformatics* **2006,** *22* (4), 400-6.

212. Sachidanandam, R.; Weissman, D.; Schmidt, S. C.; Kakol, J. M.; Stein, L. D.; Marth, G.; Sherry, S.; Mullikin, J. C.; Mortimore, B. J.; Willey, D. L.; Hunt, S. E.; Cole, C. G.; Coggill, P. C.; Rice, C. M.; Ning, Z.; Rogers, J.; Bentley, D. R.; Kwok, P. Y.; Mardis, E. R.; Yeh, R. T.; Schultz, B.; Cook, L.; Davenport, R.; Dante, M.; Fulton, L.; Hillier, L.; Waterston, R. H.; McPherson, J. D.; Gilman, B.; Schaffner, S.; Van Etten, W. J.; Reich, D.; Higgins, J.; Daly, M. J.; Blumenstiel, B.; Baldwin, J.; Stange-Thomann, N.; Zody, M. C.; Linton, L.; Lander, E. S.; Altshuler, D., A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **2001,** *409* (6822), 928-33.

213. Louie, G. V.; Bowman, M. E.; Moffitt, M. C.; Baiga, T. J.; Moore, B. S.; Noel, J. P., Structural determinants and modulation of substrate specificity in phenylalanine-tyrosine ammonia-lyases. *Chem Biol* **2006,** *13* (12), 1327-38.

214. Howles, P. A.; Sewalt, V. J. H.; Paiva, N. L.; Elkind, Y.; Bate, N. J.; Lamb, C.; Dixon, R. A., Overexpression of L-phenylalanine ammonia-lyase in transgenic tobacco plants reveals control points for flux into phenylpropanoid biosynthesis. *Plant Physiology* **1996,** *112* (4), 1617-1624.

215. Ahrens, C. H.; Brunner, E.; Qeli, E.; Basler, K.; Aebersold, R., Generating and navigating proteome maps using mass spectrometry. *Nat Rev Mol Cell Biol* **2010,** *11* (11), 789-801.

216. Shuford, C. M.; Li, Q.; Sun, Y. H.; Chen, H. C.; Wang, J.; Shi, R.; Sederoff, R. R.; Chiang, V. L.; Muddiman, D. C., Comprehensive Quantification of Monolignol-Pathway Enzymes in Populus trichocarpa by Protein Cleavage Isotope Dilution Mass Spectrometry. *J Proteome Res* **2012**.

217. Abraham, P.; Adams, R.; Giannone, R. J.; Kalluri, U.; Ranjan, P.; Erickson, B.; Shah, M.; Tuskan, G. A.; Hettich, R. L., Defining the boundaries and characterizing the landscape of functional genome expression in vascular tissues of Populus using shotgun proteomics. *J Proteome Res* **2012,** *11* (1), 449-60.

218. Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M., Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal Chem* **1996,** *68* (5), 850-8.

219. Botelho, D.; Wall, M. J.; Vieira, D. B.; Fitzsimmons, S.; Liu, F.; Doucette, A., Top-down and bottom-up proteomics of SDS-containing solutions following mass-based separation. *J Proteome Res* **2010,** *9* (6), 2863-70.

220. Wang, W.; Tai, F. J.; Chen, S. N., Optimizing protein extraction from plant tissues for enhanced proteomics analysis. *Journal of Separation Science* **2008,** *31* (11), 2032-2039.

221. Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M., Universal sample preparation method for proteome analysis. *Nat Methods* **2009,** *6* (5), 359-62.

222. Corthals, G. L.; Wasinger, V. C.; Hochstrasser, D. F.; Sanchez, J. C., The dynamic range of protein expression: a challenge for proteomic research. *Electrophoresis* **2000,** *21* (6), 1104-15.

223. Peck, S. C., Update on proteomics in Arabidopsis. Where do we go from here? *Plant Physiol* **2005,** *138* (2), 591-9.

224. Piques, M.; Schulze, W. X.; Hohne, M.; Usadel, B.; Gibon, Y.; Rohwer, J.; Stitt, M., Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in Arabidopsis. *Mol Syst Biol* **2009,** *5*, 314.

225. Second, T. P.; Blethrow, J. D.; Schwartz, J. C.; Merrihew, G. E.; MacCoss, M. J.; Swaney, D. L.; Russell, J. D.; Coon, J. J.; Zabrouskov, V., Dual-pressure linear ion trap mass spectrometer improving the analysis of complex protein mixtures. *Anal Chem* **2009,** *81* (18), 7757-65.

226. Nagaraj, N.; Kulak, N. A.; Cox, J.; Neuhauser, N.; Mayr, K.; Hoerning, O.; Vorm, O.; Mann, M., System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol Cell Proteomics* **2012,** *11* (3), M111 013722.

227. Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L., Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **2001,** *305* (3), 567-80.

228. Kuster, B.; Schirle, M.; Mallick, P.; Aebersold, R., Scoring proteomes with proteotypic peptide probes. *Nat Rev Mol Cell Biol* **2005,** *6* (7), 577-83.

229. Thakur, S. S.; Geiger, T.; Chatterjee, B.; Bandilla, P.; Frohlich, F.; Cox, J.; Mann, M., Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol Cell Proteomics* **2011,** *10* (8), M110 003699.

230. Yamada, T.; Letunic, I.; Okuda, S.; Kanehisa, M.; Bork, P., iPath2.0: interactive pathway explorer. *Nucleic Acids Res* **2011,** *39* (Web Server issue), W412-5.

231. Kruger, N. J.; Ratcliffe, R. G., Pathways and fluxes: exploring the plant metabolic network. *J Exp Bot* **2012,** *63* (6), 2243-6.

232. Punta, M.; Coggill, P. C.; Eberhardt, R. Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; Heger, A.; Holm, L.; Sonnhammer, E. L.; Eddy, S. R.; Bateman, A.; Finn, R. D., The Pfam protein families database. *Nucleic Acids Res* **2012,** *40* (Database issue), D290-301.

233. Siemens, J.; Keller, I.; Sarx, J.; Kunz, S.; Schuller, A.; Nagel, W.; Schmulling, T.; Parniske, M.; Ludwig-Muller, J., Transcriptome analysis of Arabidopsis clubroots indicate a key role for cytokinins in disease development. *Mol Plant Microbe Interact* **2006,** *19* (5), 480-94.

234. Ravanel, S.; Gakiere, B.; Job, D.; Douce, R., The specific features of methionine biosynthesis and metabolism in plants. *Proc Natl Acad Sci U S A* **1998,** *95* (13), 7805-12.

235. Guelette, B. S.; Benning, U. F.; Hoffmann-Benning, S., Identification of lipids and lipid-binding proteins in phloem exudates from Arabidopsis thaliana. *J Exp Bot* **2012,** *63* (10), 3603-16.

236. Wen, J.; Vanek-Krebitz, M.; Hoffmann-Sommergruber, K.; Scheiner, O.; Breiteneder, H., The potential of Betv1 homologues, a nuclear multigene family, as phylogenetic markers in flowering plants. *Mol Phylogenet Evol* **1997,** *8* (3), 317-33.

237. Fujimoto, Y.; Nagata, R.; Fukasawa, H.; Yano, K.; Azuma, M.; Iida, A.; Sugimoto, S.; Shudo, K.; Hashimoto, Y., Purification and cDNA cloning of cytokinin-specific binding protein from mung bean (Vigna radiata). *Eur J Biochem* **1998,** *258* (2), 794-802.

238. Kleczkowski, L. A.; Geisler, M.; Ciereszko, I.; Johansson, H., UDP-glucose pyrophosphorylase. An old protein with new tricks. *Plant Physiol* **2004,** *134* (3), 912-8.

239. Cseke, L. J.; Ravinder, N.; Pandey, A. K.; Podila, G. K., Identification of PTM5 protein interaction partners, a MADS-box gene involved in aspen tree vegetative development. *Gene* **2007,** *391* (1-2), 209-22.

240. Bohler, S.; Sergeant, K.; Lefevre, I.; Jolivet, Y.; Hoffmann, L.; Renaut, J.; Dizengremel, P.; Hausman, J. F., Differential impact of chronic ozone exposure on expanding and fully expanded poplar leaves. *Tree Physiol* **2010,** *30* (11), 1415-32.

241. Broun, P.; Poindexter, P.; Osborne, E.; Jiang, C. Z.; Riechmann, J. L., WIN1, a transcriptional activator of epidermal wax accumulation in Arabidopsis. *Proc Natl Acad Sci U S A* **2004,** *101* (13), 4706-11.

242. Volokita, M.; Rosilio-Brami, T.; Rivkin, N.; Zik, M., Combining comparative sequence and genomic data to ascertain phylogenetic relationships and explore the evolution of the large GDSL-lipase family in land plants. *Mol Biol Evol* **2011,** *28* (1), 551-65.

243. Nawrath, C., Unraveling the complex network of cuticular structure and function. *Curr Opin Plant Biol* **2006,** *9* (3), 281-7.

244. Reina, J. J.; Guerrero, C.; Heredia, A., Isolation, characterization, and localization of AgaSGNH cDNA: a new SGNH-motif plant hydrolase specific to Agave americana L. leaf epidermis. *J Exp Bot* **2007,** *58* (11), 2717-31.

245. Kurdyukov, S.; Faust, A.; Nawrath, C.; Bar, S.; Voisin, D.; Efremova, N.; Franke, R.; Schreiber, L.; Saedler, H.; Metraux, J. P.; Yephremov, A., The epidermis-specific extracellular BODYGUARD controls cuticle development and morphogenesis in Arabidopsis. *Plant Cell* **2006,** *18* (2), 321-39.

246. Peterhansel, C.; Maurino, V. G., Photorespiration redesigned. *Plant Physiol* **2011,** *155* (1), 49-55.

247. Baier, M.; Dietz, K. J., The plant 2-Cys peroxiredoxin BAS1 is a nuclear-encoded chloroplast protein: its expressional regulation, phylogenetic origin, and implications for its specific physiological function in plants. *Plant J* **1997,** *12* (1), 179-90.

248. Baier, M.; Dietz, K. J., Protective function of chloroplast 2-cysteine peroxiredoxin in photosynthesis. Evidence from transgenic Arabidopsis. *Plant Physiol* **1999,** *119* (4), 1407-14.

249. Havaux, M., Carotenoids as membrane stabilizers in chloroplasts. *Trends in Plant Science* **1998,** *3* (4), 147-151.

250. Niyogi, K. K., Photoprotection revisited: Genetic and molecular approaches. *Annu Rev Plant Phys* **1999,** *50*, 333-359.

251. Wullschleger, S. D.; Tuskan, G. A.; DiFazio, S. P., Genomics and the tree physiologist. *Tree Physiol* **2002,** *22* (18), 1273-6.

252. Bunger, M. K.; Cargile, B. J.; Sevinsky, J. R.; Stephenson, J. L., Detection and validation of coding non-synonymous SNPs from orthogonal analysis of shotgun proteomics data. *Molecular & Cellular Proteomics* **2006,** *5* (10), S367-S367.

253. Lu, B.; Xu, T.; Park, S. K.; McClatchy, D. B.; Liao, L.; Yates, J. R., 3rd, Shotgun protein identification and quantification by mass spectrometry in neuroproteomics. *Methods Mol Biol* **2009,** *566*, 229-59.

254. Searle, B. C.; Dasari, S.; Wilmarth, P. A.; Turner, M.; Reddy, A. P.; David, L. L.; Nagalla, S. R., Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm. *J Proteome Res* **2005,** *4* (2), 546-54.

255. Tsur, D.; Tanner, S.; Zandi, E.; Bafna, V.; Pevzner, P. A., Identification of post-translational modifications by blind search of mass spectra. *Nat Biotechnol* **2005,** *23* (12), 1562-7.

256. Kapp, E.; Schutz, F., Overview of tandem mass spectrometry (MS/MS) database search algorithms. *Current protocols in protein science / editorial board, John E. Coligan ... [et al.]* **2007,** *Chapter 25,* Unit25 2.

257. Sunyaev, S.; Liska, A. J.; Golod, A.; Shevchenko, A., MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal Chem* **2003,** *75* (6), 1307-15.

258. Shilov, I. V.; Seymour, S. L.; Patel, A. A.; Loboda, A.; Tang, W. H.; Keating, S. P.; Hunter, C. L.; Nuwaysir, L. M.; Schaeffer, D. A., The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol Cell Proteomics* **2007,** *6* (9), 1638-55.

259. MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R., 3rd, Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc Natl Acad Sci U S A* **2002,** *99* (12), 7900-5.

260. Choudhary, G.; Wu, S. L.; Shieh, P.; Hancock, W. S., Multiple enzymatic digestion for enhanced sequence coverage of proteins in complex proteomic mixtures using capillary LC with ion trap MS/MS. *J Proteome Res* **2003,** *2* (1), 59-67.

261. Aguiar, M.; Haas, W.; Beausoleil, S. A.; Rush, J.; Gygi, S. P., Gas-phase rearrangements do not affect site localization reliability in phosphoproteomics data sets. *J Proteome Res* **2010,** *9* (6), 3103-7.

262. Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P., A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* **2006,** *24* (10), 1285-92.

263. Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M., Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **2006,** *127* (3), 635-648.

264. Bailey, C. M.; Sweet, S. M. M.; Cunningham, D. L.; Zeller, M.; Heath, J. K.; Cooper, H. J., SLoMo: Automated Site Localization of Modifications from ETD/ECD Mass Spectra. *Journal of Proteome Research* **2009,** *8* (4), 1965-1971.

265. Chen, Y.; Chen, W.; Cobb, M. H.; Zhao, Y. M., PTMap-A sequence alignment software for unrestricted, accurate, and full-spectrum identification of post-translational modification sites. *P Natl Acad Sci USA* **2009,** *106* (3), 761-766.

266. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M., Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* **2007,** *4* (9), 709-12.

267. Michalski, A.; Neuhauser, N.; Cox, J.; Mann, M., A Systematic Investigation into the Nature of Tryptic HCD Spectra. *Journal of Proteome Research* **2012,** *11* (11), 5479-5491.

268. Reid, G. E.; Roberts, K. D.; Kapp, E. A.; Simpson, R. J., Statistical and mechanistic approaches to understanding the gas-phase fragmentation behavior of methionine sulfoxide containing peptides. *Journal of Proteome Research* **2004,** *3* (4), 751-759.

269. Ning, K.; Nesvizhskii, A. I., The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. *BMC Bioinformatics* **2010,** *11 Suppl 11*, S14.

270. Lan, P.; Li, W.; Schmidt, W., Complementary proteome and transcriptome profiling in phosphate-deficient Arabidopsis roots reveals multiple levels of gene regulation. *Mol Cell Proteomics* **2012,** *11* (11), 1156-66.

271. Wang, X. J.; Slebos, R. J. C.; Wang, D.; Halvey, P. J.; Tabb, D. L.; Liebler, D. C.; Zhang, B., Protein Identification Using Customized Protein Sequence Databases Derived from RNA-Seq Data. *Journal of Proteome Research* **2012,** *11* (2), 1009-1017.

272. Trapnell, C.; Pachter, L.; Salzberg, S. L., TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **2009,** *25* (9), 1105-1111.

273. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Proc, G. P. D., The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009,** *25* (16), 2078-2079.

274. Mortazavi, A.; Williams, B. A.; Mccue, K.; Schaeffer, L.; Wold, B., Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **2008,** *5* (7), 621-628.

275. Wright, S., Genetical structure of populations. *Nature* **1950,** *166* (4215), 247-9.

276. Bromberg, Y.; Rost, B., SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res* **2007,** *35* (11), 3823-35.

277. Rost, B.; Sander, C., Conservation and prediction of solvent accessibility in protein families. *Proteins* **1994,** *20* (3), 216-26.

278. Schlessinger, A.; Yachdav, G.; Rost, B., PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics* **2006,** *22* (7), 891-3.

279. Ng, P. C.; Henikoff, S., SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **2003,** *31* (13), 3812-4.

280. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997,** *25* (17), 3389-402.

281. Sunyaev, S. R.; Eisenhaber, F.; Rodchenkov, I. V.; Eisenhaber, B.; Tumanyan, V. G.; Kuznetsov, E. N., PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein engineering* **1999,** *12* (5), 387-94.

282. Wadskog, I.; Forsmark, A.; Rossi, G.; Konopka, C.; Oyen, M.; Goksor, M.; Ronne, H.; Brennwald, P.; Adler, L., The yeast tumor suppressor homologue Sro7p is required for targeting of the sodium pumping ATPase to the cell surface. *Molecular Biology of the Cell* **2006,** *17* (12), 4988-5003.

283. Cuff, J. A.; Clamp, M. E.; Siddiqui, A. S.; Finlay, M.; Barton, G. J., JPred: a consensus secondary structure prediction server. *Bioinformatics* **1998,** *14* (10), 892-893.

VITA

Paul Abraham was born in Madison, Wisconsin. He graduated from Bearden High School in 2003. He was introduced to the fields of mass spectrometry and proteomics by participating in undergraduate research in Robert Hettich's group at the Oak Ridge National Laboratory. Following the completion of his B.S. in 2003, he enrolled in the graduate school of Genome Science and Technology (GST) at the University of Tennessee, where he studied biological mass spectrometry and is applications in qualitative and quantitative proteomics. He expects to receive his Ph.D. in May of 2013.